DOKUZ EYLÜL UNIVERSITY

GRADUATE SCHOOL OF NATURAL AND APPLIED SCIENCES

A DATA MINING APPLICATION ON COGNITIVE EEG RECORDING

by

Alper VAHAPLAR

March, 2012

İZMİR

A DATA MINING APPLICATION ON COGNITIVE EEG RECORDING

A Thesis Submitted to the
Graduate School of Natural And Applied Sciences of Dokuz Eylül University
In Partial Fulfillment of the Requirements for the Degree of Doctor of Philosophy in Statistics

by

Alper VAHAPLAR

March, 2012
İZMİR

Ph.D. THESIS EXAMINATION RESULT FORM

We have read the thesis entitled "**A DATA MINING APPLICATION ON COGNITIVE EEG RECORDING**" completed by **ALPER VAHAPLAR** under supervision of **PROF. DR. C. CENGİZ ÇELİKOĞLU** and **PROF. DR. MURAT ÖZGÖREN** and we certify that in our opinion it is fully adequate, in scope and in quality, as a thesis for the degree of Doctor of Philosophy.

....

Prof. Dr. C. Cengiz ÇELİKOĞLU

Supervisor

....	....
Prof. Dr. Efendi NASİBOĞLU	Asst. Prof. Dr. Şen ÇAKIR
Thesis Committee Member	Thesis Committee Member
....	....
Prof. Dr. Murat ÖZGÖREN	Prof. Dr. Osman SAKA
Co-supervisor	Examining Committee Member
....	....
Assoc. Prof. Dr. Aylin ALIN	Asst. Prof. Dr. Emel KURUOĞLU
Examining Committee Member	Examining Committee Member

Prof. Dr. Mustafa SABUNCU
Director
Graduate School of Natural and Applied Sciences

ACKNOWLEDGEMENTS

Preparing a Ph. D. thesis about *Data Mining* on *Electroencephalography* was a great challenge for me at the beginning of this study. I knew it would be difficult to interfere with many topics in different domains. As I always expressed, I felt myself like a small fish (Nemo - the mascot of my thesis) in a huge ocean. I had a mountain of data which I was not familiar with, many methods and techniques which I had not known in detail and a limited period of time. When I study a subject, I saw the deepness of the ocean and I often had to take deep breaths and go deeper fearlessly(!).

I would like to express my deepest gratitude to my Supervisors, Prof. Dr. C. Cengiz ÇELİKOĞLU and Prof. Dr. Murat ÖZGÖREN for their encouragement during this journey. They always guided me to the correct path and made me feel I am not alone. They shared their experience and knowledge with me to build this study and have great contributions to my current situation. I also owe my thanks to Prof. Dr. Efendi NASİBOĞLU who excited me in my studies. The greatest contribution of my thesis emerged from the brilliant idea of him.

I would like to present my thanks especially to Prof. Dr. C. Cengiz ÇELİKOĞLU - head of Statistics Department - again for his patience and comprehence. He never refused me in even tough and busy times of him. I feel lucky of him being my advisor.

I should also thank to Dokuz Eylül University of supporting this project by maintaining especially technical instruments like computer, scanner, software and books by the project 2007.KB.FEN.003. I also would like to thank to my collaborates working in Department of Biophysics because of their help in EEG studies. The data studied in this thesis is obtained by the support of the project 2006.KB.SAG.17-38 of DEU.

My colleague in Department of Mathematics, Dr. Celal Cem SARIOĞLU, has great contribution in designing this report. So I have to thank him specially for his help and support in writing in LaTeX.

I am also grateful to my parents; my mother Melek VAHAPLAR and my father Ekrem VAHAPLAR as they always supported me in this journey. They always encouraged me in my entire educational life.

I want to express my loving thanks to my dear wife Dr. Senem ŞAHAN VAHAPLAR, who glows like sun to my life. Besides being a perfect wife, she is a wonderful collaborate and excellent researcher. She helped me to feel her assistance and encouragement always with me. She carries the meaning of *life partner* properly for me that I know she will always be with me in my entire life.

Lastly, other shining precious of mine, our daughter Duru VAHAPLAR deserves my gratitudes. She affiliated a new meaning to our lives with my wife. She is the joy of our home. Thank God for his wonderful gift.

Alper VAHAPLAR
March, 2012

A DATA MINING APPLICATION ON COGNITIVE EEG RECORDING

ABSTRACT

Development of computer and data-storage technology caused new techniques to arise to get these data useful in daily life. Especially complex statistical methods became easily usable on large amounts of data. This new approach (named as Knowledge Discovery in Databases or Data Mining) came with many advantages for every domain. It provided the transition from data to knowledge.

Human body is a complex system with subsystems in itself generating many data in various types. Brain is individually one of the vital parts of human body. It has complex communication mechanisms and many unexplored regions and functions. Electroencephalography (EEG) is a method which is used to present the electrical activity of the brain. In EEG technique, electrodes located on head receives small voltage changes produced by brain over time during a process or even in asleep. These data are used for many areas in especially epilepsy, sleep disorders, biophysics, neuroscience, etc.

This thesis aims applying some of the data mining methods on EEG data recorded during dichotic listening test. EEG data were examined in detail, analysed, partitioned and labelled. Statistical similarity measure ZM statistic was used as a tool for comparing the similarity or dissimilarity of signals received from different electrodes for different dichotic stimuli.

ZM statistic is a powerful tool in identifying similarity of signals in amplitude but not in shape. To avoid this deficiency data were transformed into difference signals to detect the behavioural similarity. Applying ZM to this transformed signals gave more reliable results in similarity. Some of the similarities which were not found before transformation arose in the transformed signals similarity. By this adjustment of data, signals moving together in different amplitudes were also detected.

Besides, a clustering was performed on electrodes using dendrogram visualization to support the similarity results.

Keywords : Data mining, electroencephalography (EEG), dichotic listening, signal similarity, ZM statistics, biomedical signals.

BİLİŞSEL EEG KAYITLARI ÜZERİNDE VERİ MADENCİLİĞİ UYGULAMASI

ÖZ

Bilgisayar ve veri saklama teknolojilerinin gelişmesi veriyi günlük hayatta daha kullanışlı hale getirmek için yeni tekniklerin ortaya çıkmasına sebep olmuştur. Özellikle karmaşık istatistiksel yöntemler, büyük miktarlardaki veriler üzerinde daha kolay uygulanabilir hale gelmiştir. Veri Tabanlarında Bilgi Keşfi ya da Veri Madenciliği isimli bu yeni yaklaşım her alana birçok avantaj getirmiştir. Bu sayede veriden tecrübeye geçiş sağlanmıştır.

İnsan vücudu kendi içinde alt sistemleri olan ve çeşitli türlerde veriler üreten bir sistemler bütünüdür. Beyin başlı başına önemli hayati organlardan birisidir. Karmaşık iletişim mekanizmalarına ve henüz keşfedilmemiş birçok bölgeye ve işleve sahiptir. Elektroensefalografi (EEG) beyindeki elektriksel aktivitenin görüntülendiği bir yöntemdir. EEG tekniğinde, kafa üzerine yerleştirilen bir başlıktaki potansiyel fark alıcıları (elektrotlar), beynin bir işlevi ya da uyku sırasında üretilen küçük voltaj değişikliklerini zaman üzerine kaydederler. Bu veriler epilepsi, uyku bozuklukları, biyofizik, nöroloji başta olmak üzere birçok alanda kullanılmaktadır.

Bu tez, veri madenciliği yöntemlerinin bazılarını dikotik dinleme testi sırasında kaydedilen EEG verileri üzerinde uygulamayı hedeflemektedir. EEG verisi detaylı olarak incelenmiş, analiz edilmiş, parçalara ayrılmış ve etiketlendirilmiştir. Farklı uyaranların etkisiyle oluşan tepkileri ve farklı elektrotlardaki sinyalleri karşılaştırmak ve benzerlik ya da benzemezliği tespit etmek üzere ZM istatistiği temel araç olarak kullanılmıştır.

ZM istatistiği sinyallerin şiddet benzerliğini belirlemede güçlü bir araç olmasına karşın şekil benzerliğini tespit etmede güçlü değildir. Tezde bu eksikliği gidermek amacıyla sinyallerin davranış benzerliğini de bulabilmek için veriler fark sinyallerine dönüştürülmüştür. ZM istatistiğini dönüştürülen verilere uygulayarak daha güvenilir sonuçlar elde edilmiştir. Dönüşümden önce bulunamayan benzerlikler fark edilir olmuştur. Verinin bu şekilde düzenlenmesiyle farklı büyüklüklerde benzer davranış gösteren sinyaller de belirlenebilmektedir.

Bunun yanısıra, elde edilen benzerliği desteklemek amacıyla elektrotlar arasında bir kümeleme çalışması da gerçekleştirilmiş ve dendrogram grafiği ile sunulmuştur.

Anahtar Sözcükler : Veri madenciliği, elektroensefalografi (EEG), dikotik dinleme, sinyal benzerliği, ZM istatistiği, biyomedikal işaretler.

CONTENTS

Page

CHAPTER ONE
INTRODUCTION

1.1 Data Mining

As a result of developing technology, decrease in cost of data storage devices, increase of data sources and getting easier to share and access any type of data, huge amounts of data has become accessible to many users in many domains. These have caused to arise environments which are rich in data but poor in data quality and knowledge. In today's competitor media of business, the importance of knowledge has been realized so the need for using the present data better for future prediction has emerged. Traditional statistical methods have been supported by faster processor and computing structures, new techniques for data processing have been developed and eventually the concept of "Data Mining" which aims to use the data to make prediction for the decision makers has been born.

Data mining is the process of applying statistical methods and analysis to huge amount of data sources in order to extract previously unknown, usable, interesting and valid information. Data mining is a step of "Knowledge Discovery in Databases (KDD)" process. In this process, methods for describing, cleaning and transforming the data, building different models for analysis, identifying the accuracy of the models and deploying the models are used.

1.2 Biomedical Signals and EEG

Living organisms are made up of many component systems - the human body, for example include the nervous system, the cardiovascular system and the musculoskeletal system, among others. Physiological processes are complex phenomena, including nervous or hormonal stimulation and control; inputs and outputs that could be in the form of physical material, neurotransmitters, or information; and action that could be mechanical, electrical or biochemical. Most physiological processes are accompanied by or manifest themselves as *signals* that reflect their nature and activities. Such signals could be of many types, including biochemical in the form of hormones and neurotransmitters, electrical in the form of potential or current, and physical in the form of pressure or temperature (Rangayyan, 2002).

The representation of biomedical signals in electronic form facilitates computer processing and analysis of the data. But many practical difficulties are encountered in biomedical signal acquisition.

The ***electroencephalogram (EEG)*** reflects the electrical activity of the brain as recorded by placing several electrodes on the scalp. The EEG is widely used for diagnostic evaluation of various brain disorders such as determining the type and location of the activity observed during an epileptic seizure or for studying sleep disorders (Sörnmo & Laguna, 2005).

The ***dichotic listening (DL)*** paradigm is often used to assess brain asymmetries at the behavioral level. Dichotic listening means presenting two auditory stimuli simultaneously, one in each ear, and the standard experiment requires that the subject report which of the two stimuli was perceived best (Hugdahl, 2005).

1.3 Problem Definition - Targets

EEG signal produced by the human brain contain many secret messages. These messages can be discovered and may be used for diagnosis or treatment of some diseases, or early detection of some problems. Analysing these signals, many different outcomes may be achieved. One of the fruitful areas is brain mapping or localization problem. According to the EEG signal analysis, the source or location of brain functions can be detected. Some diseases and causes and/or outcomes of these diseases may be defined by EEG signals.

Human brain does not define the auditory stimuli of the same intensity received from both ears equally. There is no equilibrium of 50% from right ear and 50% from left ear. Studies show that people have a right ear advantage in the rate of 60% - 70%. This thesis studies the EEG recordings of different ear advantaged subjects taken during a dichotic listening test. The responses of brain to auditory stimuli are explored, differences and similarities of right ear and left ear responses are determined, similar responses on different electrodes are detected and reasons and effects of ear advantage have been argued keeping brain asymmetry in mind.

In the study, EEG recordings received from different subjects during dichotic listening test form the basis. EEG responses are evaluated and labelled as Right Ear Advantage and Left Ear Advantage. Similarities and differences of these two responses are investigated. Similarities between different sections/electrodes and right and left ear response averages are examined to identify the functional asymmetry and functional localization of the brain. Most similar time sections of these responses are detected using different window widths. In defining similarities, cross correlation and a new statistical measure of similarity Z_M is used. The similarity methods used in signal analysis are generally on similarity of amplitude in signals. But in EEG, similarity in shape or behaviour of EEG signal is much more valuable in size or amplitude. This deficiency of Z_M statistic is overcome by transforming the signal into a difference signal.

CHAPTER TWO
SIGNAL PROCESSING AND DATA MINING

Signal processing is one of the most complicated areas in many different domains. Signals from any generator (including human body) carry many important information about the source. Understanding and working on the signal informs the researcher about the current situation, helps to predict the future states. Analysing and watching the signals, may be helpful in detecting errors, monitoring the system, preventing and avoiding possible problems and enhancing the current system components.

Signals can be stationary or non-stationary. Stationary signals are easier to work on because they have stable properties (frequencies or amplitudes). But non-stationary signals are not observed as expected mostly. Information retrieved from non-stationary signals are more valuable so far. Extracting information and defining patterns even in non-stationary behaviour of a signal is a complicated process of signal processing.

As the technology in computing speed and data storage systems develops, dream of making analysis on huge amounts of data became true. It was very difficult to apply statistical models to hundreds or thousands of data. Samples were drawn and the conclusions were made on the results of the analysis of these samples. By the developing technology, the researchers are not afraid of the amount of the data now. Microprocessors of today can make thousands of computations, databases can answer a query with millions of records just in a few seconds. So traditional statistical methods can be applied to large amounts of data. Using more data rather than samples gives more accurate results and more reliable predictions. This improves the efficiency of decision makers in a particular field of business.

The rapid change in technology also caused the statistical methods to be evolved. New and faster algorithms were developed for known methods and new methods were introduced. Estimation and prediction became more popular and easier. Different disciplines are combined and general solutions for different problems are constructed. Multi-disciplined groups perform successful operations in many fields.

Complex statistical methods cannot be thought without a computer program. Using computers and statistical software, statistical methods can be applied to any field. The corporation of computing, database and statistics emerged a new work area named *data mining*. The main target of this new area is to describe the current data, detect hidden patterns in huge amount of data and make predictions for future decisions.

2.1 Signal Analysis

The analysis of signals (especially electrical signals) is a fundamental problem for many engineers and scientists The basic parameters of interest are often changed into electrical signals by means of transducers. Common transducers include accelerometers and load cells in mechanical work, EEG electrodes and blood pressure probes in biology and medicine, and pH and conductivity probes in chemistry. The outcomes for transforming these parameters to electrical signals are great, as many instruments are available for the analysis of electrical signals in the time and frequency domains. The powerful measurement and analysis capabilities of these instruments can lead to rapid understanding of the system under study.

In this part of the thesis, the concepts of the time and frequency domains are introduced. These two ways of looking at a problem are interchangeable; that is, no information is lost in changing from one domain to another. The advantage in working these two domains is that of a change of perspective to the current situation. By changing perspective from the time domain, the solution to difficult problems can often become quite clear in the frequency domain (Agilent, 2000).

2.1.1 The Time Domain

Time domain view is the traditional way of observing signals. The time domain is a record of events in a parameter of the system versus time. Figure 2.1 shows a simple spring-mass system where a pen is attached to the mass and a piece of paper is pulled under the pen at a constant rate. The resulting drawing is a record of the displacement of the mass versus time, *a time domain view of displacement*.

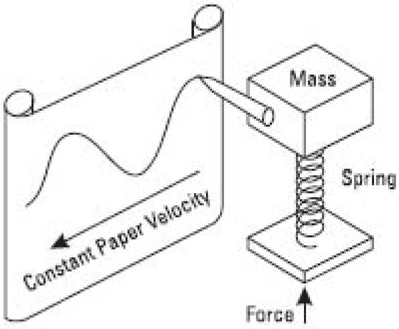


Figure 2.1 Direct recording of displacement - a time domain view (Agilent, 2000)

It is usually much more practical to convert the parameter of interest to an electrical signal using a transducer. Microphones, accelerometers, load cells, conductivity and pressure probes are the examples of transducers.

The electrical signal, which represents a parameter of the system, can be recorded on a strip chart recorder as in Figure 2.2. Doing so, the gain of the system can be adjusted to calibrate the measurement. Then the results of simple direct recording system in Figure 2.1 can be reproduced exactly.

With the indirect system a transducer can be selected which will not significantly affect the measurement by the outer effects like friction, spring and weight of the mass. This can go to the extreme of commercially available displacement transducers which do not even contact the mass. The pen deflection can be easily set to any desired value by controlling the gain of the electronic amplifiers.

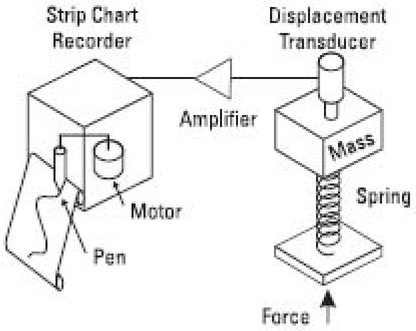


Figure 2.2 Indirect recording of displacement (Agilent, 2000)

This indirect system works well until the measured parameter begins to change rapidly. Because of the mass of the pen and recorder mechanism and the power limitations of its drive, the pen can only move at finite velocity. If the measured parameter changes faster, the output of the recorder will be in error. A common way to reduce this problem is to eliminate the pen and record on a photosensitive paper by deflecting a light beam. Such a device is called an *oscillograph*. Since it is only necessary to move a small, light-weight mirror through a very small angle, the oscillograph can respond much faster than a strip chart recorder.

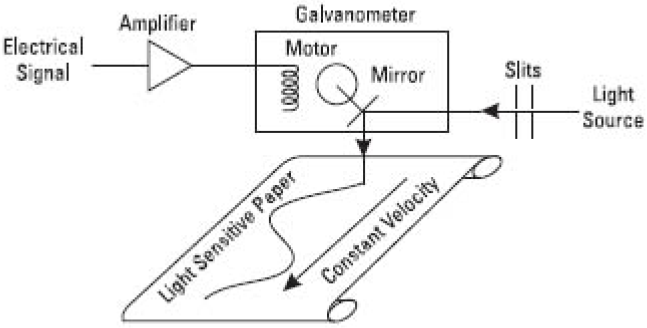


Figure 2.3 Simplified oscillograph operation (Agilent, 2000)

Another common device for displaying signals in the time domain is the *oscilloscope*. Here an electron beam is moved using electric fields. The electron beam is made visible by a screen of phosphorescent material. It is capable of accurately displaying signals that vary even more rapidly than the oscillograph can handle. This is because it is only necessary to move an electron beam, not a mirror.

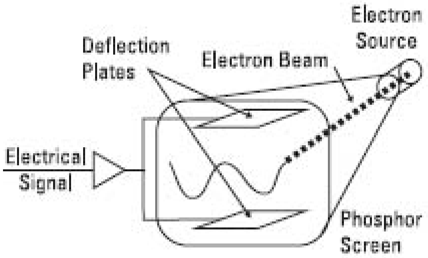


Figure 2.4 Basic oscilloscope operation (Agilent, 2000)

The strip chart, oscillograph and oscilloscope all show displacement versus time. Changes in this displacement represent the variation of the parameter versus time.

2.1.2 The Frequency Domain

It was shown over one hundred years ago by Baron Jean Baptiste Fourier that any waveform that exists in the real world can be generated by adding up sine waves. This was illustrated in Figure 2.5 for a simple waveform composed of two sine waves. By regulating the amplitudes, frequencies and phases of these sine waves correctly, a waveform can be generated identical to the desired signal. Conversely, any real world signal can be broken down into sine waves.

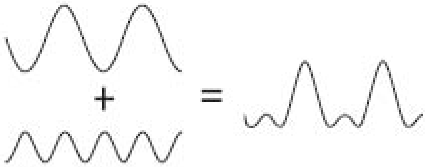


Figure 2.5 Any real waveform can be produced by adding sine waves together. (Agilent, 2000)

Figure 2.6a is a three dimensional graph of this addition of sine waves. Two of the axes are time and amplitude, familiar from the time domain. The third axis is frequency which allows to visually separate the sine waves which add to give out the complex waveform. Viewing this three-dimensional graph along the frequency axis, view in Figure 2.6b is obtained. This is the time domain view of the sine waves. Adding them together at each instant of time gives the original waveform. However, if the graph is viewed along the time axis as in Figure 2.6c, a totally different picture is displayed. Here the axes of amplitude versus frequency, is commonly called the frequency domain. Every sine wave separated from the input appears as a vertical line. Its height represents its amplitude and its position represents its frequency. Since each line represents a sine wave, the input signal is uniquely characterized in the frequency domain. This frequency domain representation of a signal is called the *spectrum* of the signal. Each sine wave line of the spectrum is called a *component* of the total signal.

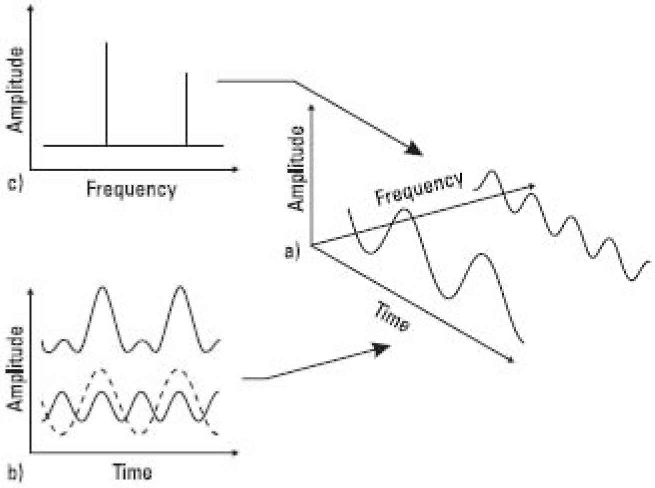


Figure 2.6 The relationship between the time and frequency domains. a) Three dimensional coordinates showing time, frequency and amplitude b) Time domain view c) Frequency domain view (Agilent, 2000)

It should be expressed that *information is neither gained nor lost, just is represented differently.* The same three-dimensional graph is viewed from different angles. This different perspective can be very useful. At first the frequency domain may seem strange and unfamiliar, yet it is an important part of everyday life. The ear-brain combination is an excellent frequency domain analyser. The ear-brain splits the audio spectrum into many narrow bands and determines the power present in each band. It can easily pick small sounds out of loud background noise thanks in part to its frequency domain capability. A doctor listens to the patient's heart and breathing for any unusual sounds. An experienced mechanic can do the same thing with a machine. Using a screwdriver as a stethoscope, he can hear when a bearing is failing because of the frequencies it produces.

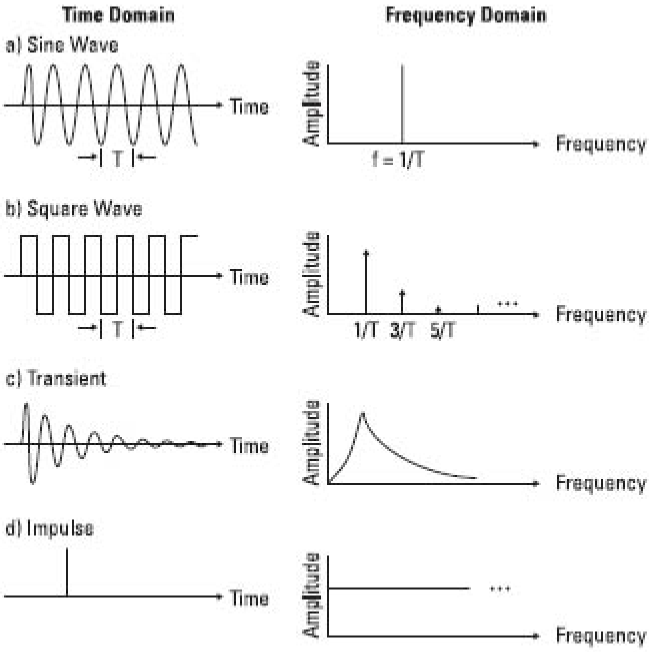


Figure 2.7 Frequency spectrum examples (Agilent, 2000)

In Figure 2.7a, it is seen that the spectrum of a sine wave is just a single line. The square wave in Figure 2.7b is made up of an infinite number of sine waves, all harmonically related. This is in contrast to the transient signal in Figure 2.7c which has a continuous spectrum. Another signal of interest is the impulse shown in Figure 2.7d in which there is energy at all frequencies.

2.2 Signal Similarity Methods

Measures of similarity are required in a wide range of radar sonar, communications, remote sensing, artificial intelligence and medical applications, where one signal or image is compared with another. Many basic signal processing operations, such as matched filtering, cross correlation, and beam formation are based on measures of similarity. These operations form the foundation of the detection, classification, localization, association and registration algorithms employed in semiautonomous sensor systems. Beam-formation and cross-correlation processing techniques are also used to compute Time-Of-Arrival- Differences (TOADs) or Time Delay Estimates (TDE) in distribu-ted networks of acoustic sensors (Kennedy, 2007).

Many signals have similarities that can be exploited in signal processing algorithms. For example, a phase-modulated signal is similar to an amplitude-scaled version of that signal; processing to extract the information should ideally be invariant to changes in amplitude. In circumstances where similarities can be identified, it may be desirable to design signal processing algorithms that are invariant to the different forms of the signal that are fundamentally similar in some aspect. Many signal processing algorithms have been developed that attempt to compensate for differences in amplitude, offset, phase, or time. However, these have all been developed separately without regard to a unifying principle (Moon, 1996).

2.2.1 Signal Transformations

Mathematical transformations are applied to signals to obtain a further information from that signal that is not readily available in the raw signal (signals in time domain).

There are number of transformations that can be applied, among which the Fourier transforms are probably by far the most popular that breaks down a signal into constituent sinusoids of different frequencies. Another way to think of Fourier analysis is as a mathematical technique for *transforming* the view of the signal from *time-based* to *frequency-based*.

Most of the signals in practice, are time domain signals in their raw format. That is, whatever that signal is measuring, is a function of time. In other words, when plotting the signal one of the axes is time (independent variable), and the other (dependent variable) is usually the amplitude. This representation is not always the best representation of the signal for most signal processing related applications. In many cases, the most distinguished information is hidden in the frequency content of the signal. The frequency SPECTRUM of a signal is basically the frequency components (spectral components) of that signal. The frequency spectrum of a signal shows what frequencies exist in the signal (Polkar, 2001).

Frequency is something to do with the change in rate of something. If something changes rapidly, we say that it is of high frequency, where as if this variable does not change rapidly, i.e., it changes smoothly, we say that it is of low frequency. If this variable does not change at all, then we say it has zero frequency, or no frequency. For example the publication frequency of a daily newspaper is higher than that of a monthly magazine (it is published more frequently).

As expressed in Polkar (2001), the frequency is measured in cycles/second, or with a more common name, in "Hertz". For example the electric power we use in our daily life 50 Hz. This means that if you try to plot the electric current, it will be a sine wave passing through the same point 50 times in 1 second. In the following figures, the first one is a sine wave at 3 Hz, the second one at 10 Hz, and the third one at 50 Hz.

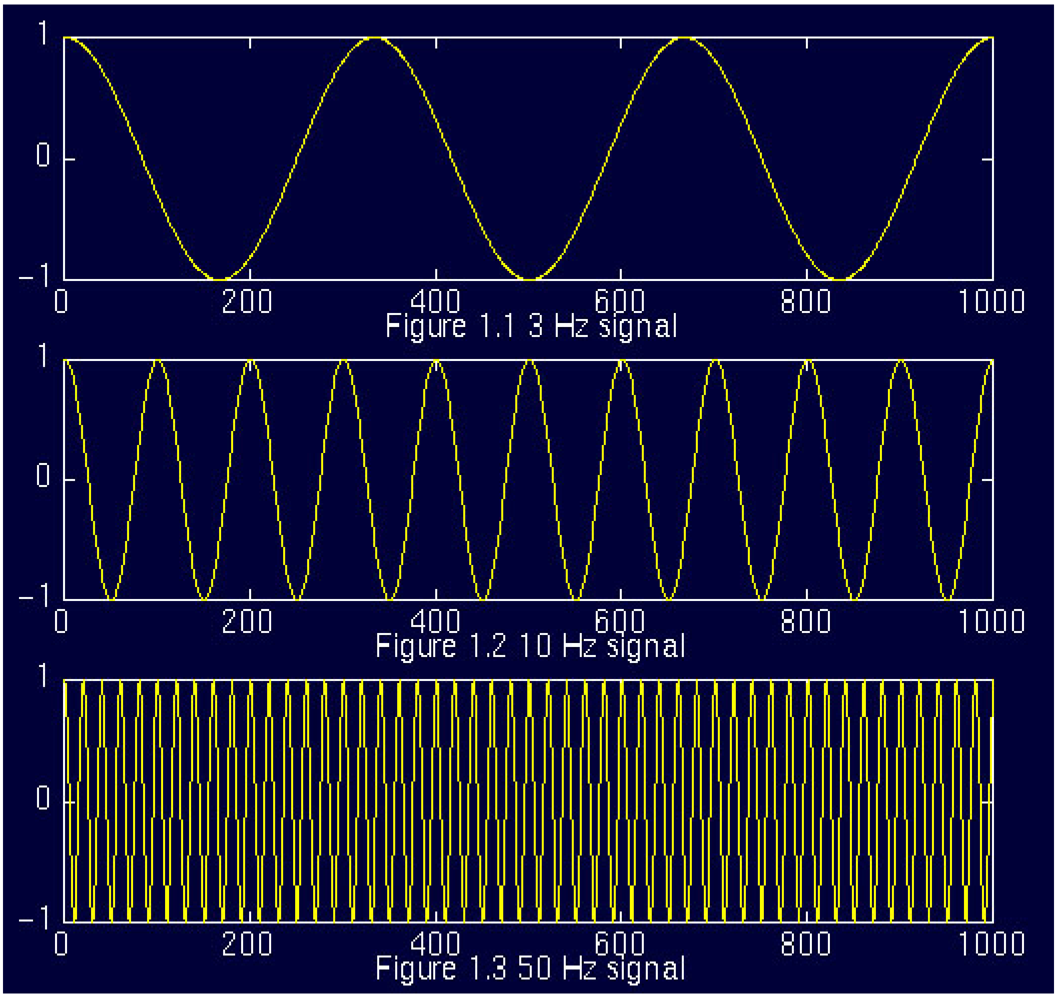


Figure 2.8 Signals in different frequencies (Polkar, 2001)

Why need to transform?

Depending on the target of the analysis, the information that cannot be readily seen in the time-domain can be seen in the frequency domain. Especially if the work is about frequencies, time domain plotting will not be helpful for the researcher.

Let's give an example from biological signals. Suppose we are looking at an ECG signal (ElectroCardioGraphy, graphical recording of heart's electrical activity). The typical shape of a healthy ECG signal is well known to cardiologists. Any significant deviation from that shape is usually considered to be a symptom of a pathological condition. This pathological condition, however, may not always be quite obvious in the original time-domain signal. Cardiologists usually use the time-domain ECG signals which are recorded on strip-charts to analyse ECG signals. Recently, the new computerized ECG recorders/analysers also utilize the frequency information to decide whether a pathological condition exists. A pathological condition can sometimes be diagnosed more easily when the frequency content of the signal is analysed (Polkar, 2001).

This, of course, is only one simple example why frequency content might be useful. Today Fourier Transforms are used in many different areas including all branches of engineering.

Although FT is probably the most popular transform being used (especially in electrical engineering), it is not the only one. There are many other transforms that are used quite often by engineers and mathematicians. Hilbert transform, short-time Fourier transform, Wigner distributions, the Radon Transform, and of course our featured transformation, the wavelet transform, constitute only a small portion of a huge list of transforms that are available at engineer's and mathematician's disposal. Every transformation technique has its own area of application, with advantages and disadvantages, and the wavelet transform (WT) is no exception. For example WT is useful when to have both the time and the frequency information at the same time.

Signals whose frequency content do not change in time are called stationary signals. In other words, the frequency content of stationary signals do not change in time. In this case, one does not need to know at what times frequency components exist , since all frequency components exist at all times.

An example of time domain to frequency domain transformation with FT is given below for example the stationary signal $x(t) = cos(2\pi 10t) + cos(2\pi 25t) + cos(2\pi 50t) + cos(2\pi 100t)$. It is stationary because it has frequencies of 10, 25, 50, and 100 Hz at any given time instant. This signal is plotted below:

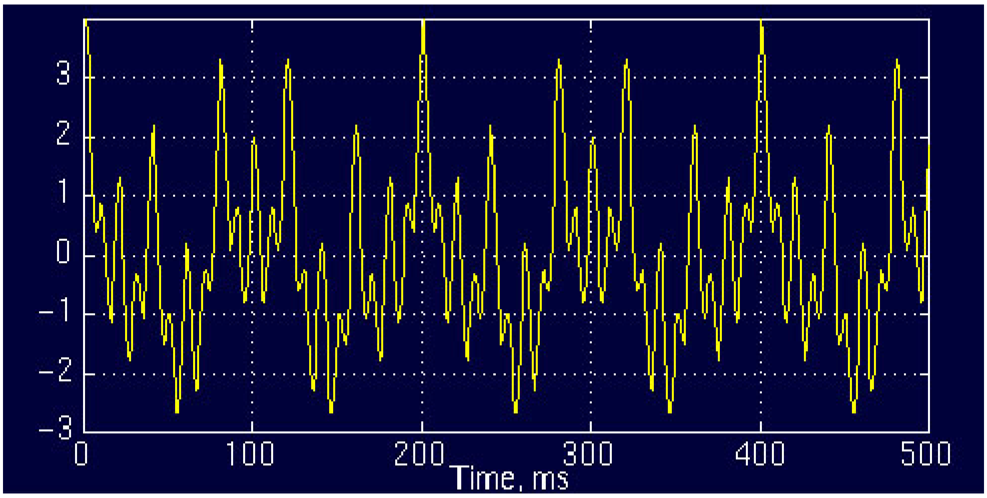


Figure 2.9 Signal of $x(t) = cos(2\pi 10t) + cos(2\pi 25t) + cos(2\pi 50t) + cos(2\pi 100t)$ (Polkar, 2001)

And the following FT is:

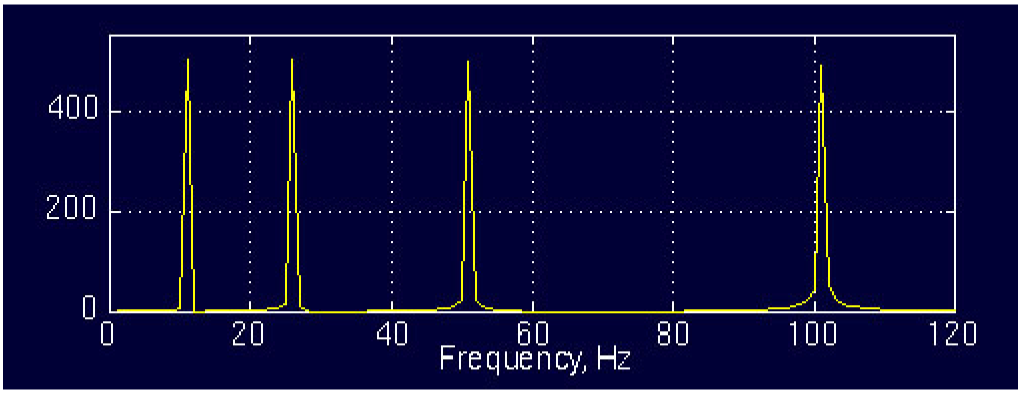


Figure 2.10 FT of $x(t) = cos(2\pi 10t) + cos(2\pi 25t) + cos(2\pi 50t) + cos(2\pi 100t)$ (Polkar, 2001)

While working on the signals, application specific transformations can also be used. Frequency or amplitude filtering, amplification, normalization or averaging are also used techniques for transformation. In this study, the signals are transformed by calculating the difference of each instance within the signal as explained in Section 4.6.1.

2.2.2 Z_M Statistic

In Kennedy (2007) a statistical treatment of a delay-and-sum beam-former is described and used to derive the new measure of signal similarity. The derivation is based on a few standard statistical relationships. A hypothesis test is performed with the null hypothesis being that there is no signal present and that the waveforms entering the beam former contain only zero-mean Gaussian-distributed noise. It is assumed that any Direct Current (DC) offset in the data (e.g. sensor bias) or frequencies that are of no interest (e.g. wind or self noise) have been removed by a pre-whitening stage. If the null hypothesis is rejected then it is assumed that a localizable signal is present. The test statistic for all possible lag combinations corresponding to all physically measurable angles is computed. The most likely direction of the source is set equal to the angular coordinate for which the null hypothesis is least likely, i.e. the test statistic is maximized.

In the study of Kennedy (2007), the delay-and-sum beam-former is applied as

$$y(n) = \sum_{m=0}^{M-1} x_m(n) \tag{2.2.1}$$

where $x_m(n)$ is the n^{th} sample output from the m^{th} delay channel and $y(n)$ is the beam-formed output. In Eq. (2.2.1) it is assumed that the appropriate delays have been applied to steer a beam in a desired direction. The noise statistics of every sample from all sensors are assumed to be identical, so the n^{th} sample in each delay channel is assumed to be an independent observation of the random variable X_n. Analysing the digitized waveforms (in x) over a window of length N, gives a total of N different random variables, with M observations of each variable. Under the null hypothesis the variables have a Gaussian (Normal) distribution

$$X_N \sim N(\mu_n, \sigma_n^2) \tag{2.2.2}$$

At a given n, using the data from all M channels, the Maximum Likelihood Estimates (MLEs) $\hat{\mu}_n$ and $\hat{\sigma}_n^2$, of the (true) mean and variance μ_n and σ_n^2 , are computed using

$$\hat{\mu}_n = \frac{y(n)}{M} \tag{2.2.3}$$

and

$$\hat{\sigma}_n^2 = \frac{1}{M}\left\{\sum_{m=0}^{M-1} x_m(n)^2 - \frac{y(n)^2}{M}\right\} \tag{2.2.4}$$

Under the null hypothesis the following relationships hold:

$$\text{If } Z_a = M\frac{(\hat{\mu}_n - \mu_n)^2}{\sigma_n^2} \text{ then } Z_a \sim \chi^2(1) \tag{2.2.5}$$

$$\text{If } Z_b = M\frac{\hat{\sigma}_n^2}{\sigma_n^2} \text{ then } Z_b \sim \chi^2(M-1) \tag{2.2.6}$$

Under the null hypothesis it is also assumed that the noise statistics of the sensor outputs are zero mean and time invariant so the parameters of each distribution are the same:

$$\mu_1 = \mu_2 = \ldots = \mu_N = \mu = 0 \tag{2.2.7}$$

and

$$\sigma_1^2 = \sigma_2^2 = \ldots = \sigma_N^2 = \sigma^2 = 0 \tag{2.2.8}$$

Using the reproductive property of χ^2 variables, the following aggregate test statistics can be formed and analyzed:

$$\text{If } Z_c = \frac{M}{\sigma^2}\sum_{n=0}^{N-1} \hat{\mu}_n^2 \text{ then } Z_c \sim \chi^2(N) \tag{2.2.9}$$

$$\text{If } Z_d = \frac{M}{\sigma^2}\sum_{n=0}^{N-1} \hat{\sigma}_n^2 \text{ then } Z_d \sim \chi^2(N(M-1)) \tag{2.2.10}$$

So far it has been assumed that the true variance (σ^2) of the (white) noise is known. This is an inconvenient and unnecessary assumption. It can be eliminated by dividing (2.2.9) by (2.2.10); furthermore, if the numerator and the denominator are scaled by the inverse of their respective degrees of freedom, i.e.

$$Z_M = \frac{Z_c/N}{Z_d/(N(M-1))} \tag{2.2.11}$$

then a variable distributed according to Snedecor's F distribution results (Freund, 1992, Kennedy, 2007); that is, after substituting (2.2.9) and (2.2.10) into (2.2.11):

$$Z_M = (M-1)\frac{\sum\limits_{n=0}^{N-1} \hat{\mu}_n^2}{\sum\limits_{n=0}^{N-1} \hat{\sigma}_n^2} \tag{2.2.12}$$

with

$$Z_M \sim F(N, N(M-1)) \tag{2.2.13}$$

$$Z_M = (M-1)\frac{\frac{1}{M}\sum\limits_{n=0}^{N-1} y(n)^2}{\sum\limits_{m=0}^{M-1}\sum\limits_{n=0}^{N-1} x_m(n)^2 - \frac{1}{M}\sum\limits_{n=0}^{N-1} y(n)^2} \tag{2.2.14}$$

Alternatively, 2.2.14 may be written in terms of moments:

$$Z_M = (M-1)\frac{\sum\limits_{n=0}^{N-1} E[x(n)]^2}{\sum\limits_{n=0}^{N-1} E[x(n)^2] - \sum\limits_{n=0}^{N-1} E[x(n)]^2} \tag{2.2.15}$$

using

$$E[x(n)] = \frac{1}{M}\sum_{m=0}^{M-1} x_m(n) \tag{2.2.16}$$

$$E[x(n)^2] = \frac{1}{M}\sum_{m=0}^{M-1} x_m(n)^2 \tag{2.2.17}$$

As expressed in Kennedy (2007) the Z_M test statistic is the ratio of two sum-of-squares quantities (2.2.12). If the square of the estimated mean (numerator) is regarded as the (delay-and-sum) signal power, and the variance (denominator) the noise power, then it may be convenient to convert Z_M into a Signal-to-Noise Ratio (SNR) in dB. Images may then be formed using many closely-spaced beams, and presented to an operator for visual inspection.

Knowing that under the null hypothesis the Z_M statistic is F distributed, allows a detection threshold to be computed to give the desired probability of falsely rejecting the null hypothesis when it is indeed true (a false alarm). If the computed Z_M value exceeds the threshold then a localisable signal is instead assumed to be present. The necessary threshold is determined using the inverse Cumulative Density Function (CDF) of the F distribution. The two parameters (degrees of freedom) of the function automatically adjust the threshold (increase it) to compensate for the higher variability of the test statistic when low channel counts (M) are used and when the data window length (N) is small.

In practice, the null hypothesis is rarely entirely true, and false alarms due to nuisance sources are common, so a larger detection threshold is usually appropriate, giving a negligible theoretical false-alarm probability (the size of the test), an acceptable practical false-alarm probability and a reasonable probability of detection (the power of the test) (Kennedy, 2007).

2.3 Data Mining

Due to the emerging data storages in databases, the lack of information and knowledge arises in every field of daily life. As early as 1984, in his book Megatrends, John Naisbitt observed that “we are drowning in information but starved for knowledge.” The problem today is not that there is not enough data and information streaming in. We are, in fact, inundated with data in most fields. Rather, the problem is that there are not enough trained human analysts available who are skilled at translating all of these data into knowledge.

We are overwhelmed with data. The amount of data in the world, in our lives, seems to go on and on increasing and there’s no end in sight. Personal computers make it too easy to save things that previously we would have trashed. Inexpensive multi gigabyte disks make it too easy to postpone decisions about what to do with all this stuff we simply buy another disk and keep it all. Different types of electronic equipments record our decisions, our choices in the supermarket, our financial habits, our comings and goings. We swipe our way through the world, every swipe a record in a database. The World Wide Web overwhelms us with information; meanwhile, every choice we make is recorded. And all these are just personal choices: they have countless counterparts in the world of commerce and industry. We would all testify to the growing gap between the generation of data and our understanding of it. As the volume of data increases, inexorably, the proportion of it that people understand decreases, alarmingly. Lying hidden in all these data is information, potentially useful information, that is rarely made explicit or taken advantage of (Witten & Frank, 2005).

The steady and amazing progress of computer hardware technology in the past three decades has led to powerful, affordable, and large supplies of computers, data collection equipment, and storage media. This technology provides a great boost to the database and information industry, and makes a huge number of databases and information repositories available for transaction management, information retrieval, and data analysis (Han & Kamber, 2001).

Traditional data analysis techniques have often encountered practical difficulties in meeting the challenges posed by new data sets. The following are some of the specific challenges that motivated the development of data mining:

- Scalability,
- High dimensionality,
- Heterogeneous and complex data,
- Data ownership and distribution,
- Non-traditional analysis.

Brought together by the goal of meeting these challenges, researchers from different disciplines began to focus on developing more efficient and scalable tools that could handle diverse types of data. This work, which culminated in the field of data mining, built upon the methodology and algorithms that researchers had previously used. In particular, data mining draws upon ideas such as sampling, estimation and hypothesis testing from statistics, search algorithms, modelling techniques and learning theories from artificial intelligence, pattern recognition and machine learning. Data mining has also been quickly adopt ideas from other areas including optimization, evolutionary computing, information theory, signal processing, visualization, and information retrieval. (Tan et al., 2006)

Data can now be stored in many different types of databases. One database architecture that has recently emerged is the data warehouse, a repository of multiple heterogeneous data sources, organized under a unified schema at a single site in order to facilitate management decision making. Data warehouse technology includes data cleansing, data integration, and On-Line Analytical

Processing (OLAP), that is, analysis techniques with functionalities such as summarization, consolidation and aggregation, as well as the ability to view information at different angles. Although OLAP tools support multidimensional analysis and decision making, additional data analysis tools are required for in-depth analysis, such as data classification, clustering, and the characterization of data changes over time (Han & Kamber, 2001).

Data mining is an interdisciplinary field, the confluence of a set of disciplines (as shown in Figure 2.11), including database systems, statistics, machine learning, visualization, and information science. Moreover, depending on the data mining approach used, techniques from other disciplines may be applied, such as neural networks, fuzzy and/or rough set theory, knowledge representation, inductive logic programming, or high performance computing. Depending on the kinds of data to be mined or on the given data mining application, the data mining system may also integrate techniques from spatial data analysis, information retrieval, pattern recognition, image analysis, signal processing, computer graphics, web technology, economics, or psychology (Han & Kamber, 2001).

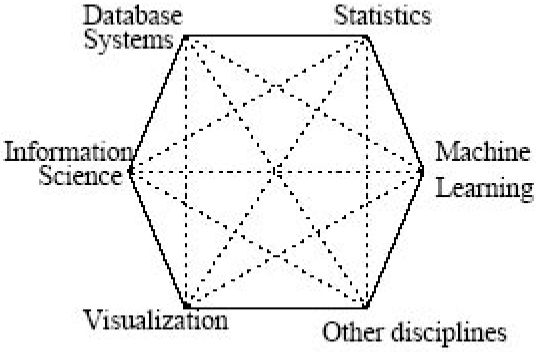


Figure 2.11 Data mining as a confluence of multiple disciplines (Han & Kamber, 2001)

2.3.1 Knowledge Discovery

Simply stated, data mining refers to extracting or "*mining*" knowledge from large amounts of data. Thus, "*data mining*" should have been more appropriately named "*knowledge mining from data*". *Knowledge mining*, a shorter term, may not reflect the emphasis on mining from large amounts of data. Nevertheless, mining is a vivid term characterizing the process that finds a small set of precious nuggets from a great deal of raw material. There are many other terms carrying a similar or slightly different meaning to data mining, such as knowledge mining from databases, knowledge extraction, data/pattern analysis, data archaeology, and data dredging (Han & Kamber, 2001, Larose, 2005, Bramer, 2007, Tan et al., 2006).

Many people treat data mining as a synonym for another popularly used term, *Knowledge Discovery in Databases*, or *KDD*. Alternatively, others view data mining as simply an essential step in the process of knowledge discovery in databases.

According to Han & Kamber (2001), KDD is a process containing the following steps:

- Data cleaning
- Data integration
- Data selection
- Data transformation
- Data mining
- Pattern evaluation
- Knowledge presentation

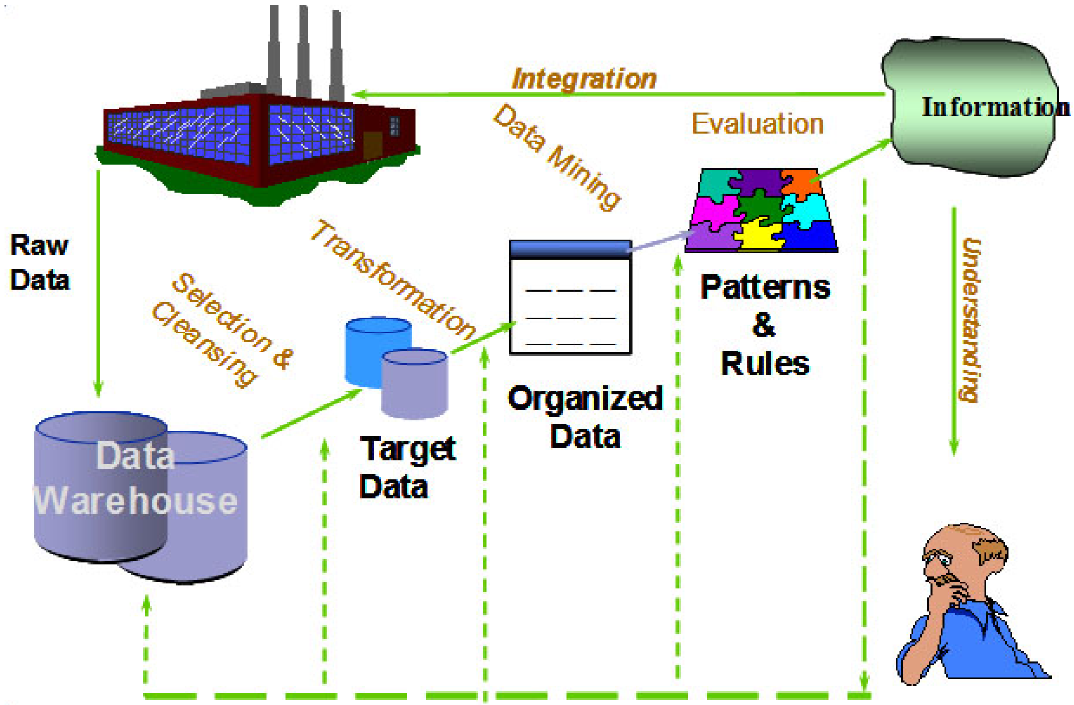


Figure 2.12 Knowledge Discovery Cycle

2.3.2 CRISP-DM Life Cycle

There is a temptation in some companies, due to departmental inertia and compart-mentalization, to approach data mining haphazardly, to reinvent the wheel and duplicate effort. A cross-industry standard was clearly required that is industry-neutral, tool-neutral, and application-neutral. The Cross-Industry Standard Process for Data Mining (CRISP-DM) was developed in 1996 by analysts representing DaimlerChrysler, SPSS, and NCR. CRISP provides a nonproprietary and freely available standard process for fitting data mining into the general problem-solving strategy of a business or research unit.

According to CRISP-DM expressed in Larose (2005), a given data mining project has a life cycle consisting of six phases, as illustrated in Figure 2.13. Note that the phase sequence is adaptive. That is, the next phase in the sequence often depends on the outcomes associated with the preceding phase. The most significant dependencies between phases are indicated by the arrows. For example, suppose that we are in the modeling phase. Depending on the behavior and characteristics of the model, we may have to return to the data preparation phase for further refinement before moving forward to the model evaluation phase.

The iterative nature of CRISP is symbolized by the outer circle in Figure 2.13. Often, the solution to a particular business or research problem leads to further questions of interest, which may then be attacked using the same general process as before.

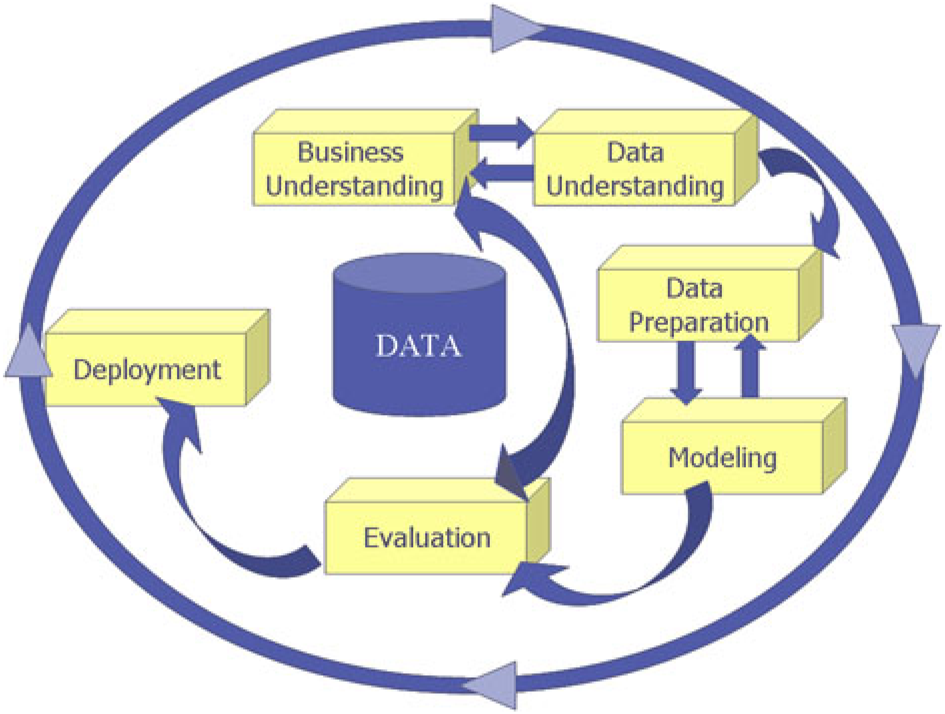


Figure 2.13 CRISP-DM Life Cycle

Lessons learned from past projects should always be brought to bear as input into new projects. Following is an outline of each phase. Although conceivably, issues encountered during the evaluation phase can send the analyst back to any of the previous phases for amelioration, for simplicity we show only the most common loop, back to the modelling phase.

In Larose (2005) the six phases of CRISP-DM are expressed as follows:

1. *Business understanding phase:* The first phase in the CRISP-DM standard process may also be termed the research understanding phase.

- Enunciate the project objectives and requirements clearly in terms of the business or research unit as a whole.
- Translate these goals and restrictions into the formulation of a data mining problem definition.
- Prepare a preliminary strategy for achieving these objectives.

2. *Data Understanding phase:*

- Collect the data
- Use exploratory data analysis to familiarize yourself with the data and discover initial insights.
- Evaluate the quality of the data.
- If desired, select interesting subsets that may contain actionable patterns.

3. *Data Preparation Phase:*

- Prepare from the initial raw data the final data set that is to be used for all subsequent phases. This phase is very labor intensive.
- Select the cases and variables you want to analyze and that are appropriate for your analysis.
- Perform transformations on certain variables, if needed.
- Clean the raw data so that it is ready for the modeling tools.

4. *Modeling Phase:*

- Select and apply appropriate modeling techniques.
- Calibrate model settings to optimize results.
- Remember that often, several different techniques may be used for the same data mining problem.
- If necessary, loop back to the data preparation phase to bring the form of the data into line with the specific requirements of a particular data mining technique.

5. *Evaluation Phase:*

- Evaluate the one or more models delivered in the modeling phase for quality and effectiveness before deploying them for use in the field.
- Determine whether the model in fact achieves the objectives set for it in the first phase.
- Establish whether some important facet of the business or research problem has not been accounted for sufficiently.
- Come to a decision regarding use of the data mining results.

6. *Deployment Phase:*

- Make use of the models created.
- Example of a simple deployment: Generate a report.
- Example of a more complex deployment: Implement a parallel data mining process in another department.
- For businesses, the customer often carries out the deployment based on your model.

2.3.3 Methods and Tasks

In the knowledge discovery process, many methods and techniques must be used according to the type of the data and the target of the study. In order to understand and describe the data, find hidden patterns, apply statistical models and to use the data for prediction, various methods must be experienced. To get reliable results, many methods and tasks must be tried.

Methods used in data mining process can be classified under two categories - *supervised* and *unsupervised*. In supervised techniques there is a target attribute and the class label of each sample is provided. In other words, learning of the model is supervised in that it is told to which class each training sample belongs. Many of the methods - especially classification methods - used in data mining are supervised (Tan et al., 2006, Bramer, 2007, Han & Kamber, 2001, Larose, 2005).

In unsupervised techniques, no target attribute exists or the class of the target is undefined before training. Also the class labels of training samples are not known. Clustering is an example of unsupervised models.

The general classification of the tasks used in knowledge discovery process are as follows (Larose, 2005):

- Description
- Clustering
- Classification
- Estimation - Prediction
- Association

2.3.3.1 Description

Data often contains many information in the first sight. Looking at the big picture and then detailing it may give fruitful outcomes for decision makers. Before going in detail, data must be described fully. Database management systems mostly offer a data dictionary for the data kept, in terms of data type, storage, a short description about the values stored, etc.

As a good representation of the data, *visualization techniques* (sometimes called graphical data analysis) provide rewarding understanding in discovering patterns or relations in the data. Using different types of charts (bar, box plot, stem and leaf, scatter plot, pie, web graphs, etc.) and tables help to see what is in a data set. Matrix plots, distribution diagrams, histograms, cross tabulations and correlations clearly define the relations between the attributes. Tools for representing data in 2, 3 and even more dimensions exists in today's technology. These tools provide different views for the data.

Besides the visual representation, some numerical values must be obtained for a better understanding. Descriptive statistics like minimum and maximum values, ranges, frequencies, averages, modes, standard deviations, variances, quartiles or deciles, cumulative percentiles, correlation coefficients are useful and simple computations for representing attributes kept in the data (Vahaplar, 2003).

As mentioned in Larose (2005), describing the data is the concern of a specific subject named *Exploratory Data Analysis* which allows the analyst to

- represent the data deeply in terms of graphical, tabular and numerical tools,
- examine the interrelations among the attributes,
- construct new subsets of data according to the related scenario or cases.

2.3.3.2 Clustering

Clustering refers to the grouping of records, observations, or cases into classes of similar objects. A cluster is a collection of records that are similar to one another and dissimilar to records in other clusters. Clustering differs from classification in that there is no target variable for clustering. The clustering task does not try to classify, estimate, or predict the value of a target variable (unsupervised). Instead, clustering algorithms seek to segment the entire dataset into relatively homogeneous subgroups or clusters, where the similarity of the records within the cluster is maximized, and the similarity to records outside this cluster is minimized.

Clustering is often performed as a preliminary step in a data mining process, with the resulting clusters being used as further inputs into a different technique downstream, such as neural networks. Due to the enormous size of many present-day databases, it is often helpful to apply clustering analysis first, to reduce the search space for the downstream algorithms. (Hartigan, 1975, Grabmaier & Rudolph, 2002)

In clustering, there are some issues to be encountered such as measuring similarity (or dissimilarity) between records, dealing with categorical variables, normalization of numerical attributes and determining the optimum number of clusters.

There are different algorithms used in clustering. Basically clustering algorithms are classified as follows: (Gan et al., 2007, Ulutagay, 2009)

- Hierarchical Clustering Methods (Connectivity based),
  Agglomerative methods, Divisive methods, (CURE, BIRCH)
- Partitioning Methods (Centroid based, center based),
  k-means, k-modes, k-medoids, k-prototypes, k-probabilities

- Density Based Methods,

  DBSCAN, GDBSCAN, OPTICS

- Searched Based Methods,

- Grid Based Methods,

  STING, CLIQUE

- Graph Based Methods,

- Fuzzy Clustering Methods,

  Fuzzy c-means (FCM), Fuzzy Joint Point (FJP)

- Model Based Methods

  COBWEB, CLASSIT, AutoClass, Kohonen Self Organizing Maps.

2.3.3.3 Classification

Classification and Prediction are two forms of data analysis which can be used to extract models describing important data classes or to predict future trends.

Data classification is a two-step process. In the first step named learning, a model is built using a set of data (*training set*) with predefined classes. The model analyses the records each belongs to a predefined class. One of the attributes in the data is called *class label attribute*. The elements of the training set is selected randomly from the sample population. The model is represented as classification rules, mathematical formulae or decision trees.

In the second step called classification, the model built is used for classification of future data which class labels are not known. According to the rules or formulae constructed in model, the class which the record must belong to is determined.

Classification techniques and some favourite algorithms are as follows (Kotsiantis, 2007, Han & Kamber, 2001, Bramer, 2007) :

- Logic Based Algoritms

  decision trees (C4.5, CART, CHAID, QUEST), learning set of rules,

- Perceptron-based techniques

  Single layered (WINNOW), multi layered (Artificial Neural Networks), Radial Basis Function (RBF) networks,

- Statistical Learning Algorithms

  Naive Bayes classifiers, Bayesian networks,

- Instance Based Learning

  k-Nearest Neighbour (kNN),

- Case Based Reasoning,

- Support Vector Machines,

- Genetic Algorithms,

- Rough Set Approach,

- Fuzzy Set Approach.

2.3.3.4 Estimation - Prediction

Estimation is similar to classification except that the target variable is numerical rather than categorical. Models are built using “complete” records, which provide the value of the target variable as well as the predictors. Then, for new observations, estimates of the value of the target variable are made, based on the values of the predictors.

Prediction is similar to classification and estimation, except that for prediction, the results lie in the future. Prediction is the constructing and use of a model to assess the class of an unlabelled sample or to assess the value or value range of an attribute that a given sample likely to have. Basically, classification and regression are two major types of prediction problems. Classification is used to predict discrete or nominal variables and regression is used to predict continuous or ordered variables. The prediction of continuous values can be modelled by statistical techniques of regression.

In statistical analysis, estimation and prediction are elements of the field of *statistical inference*. Statistical inference consists of methods for estimating and testing hypotheses about population characteristics based on the information contained in the sample. The unknown value of the population mean μ is estimated by calculating the average values of a sample $\overline{x}$ drawn from that population. The sample proportion p is the statistic used to measure the unknown value of the population proportion π. The statistic s is used to estimate the standard deviation σ of the population (Larose, 2005).

In some cases, a point estimation has to be done but in many cases, confidence interval estimation is more efficient. A *confidence interval estimate* of a population parameter consists of an interval of numbers produced by a point estimate, together with an associated confidence level specifying the probability that the interval contains the parameter and expressed as ***point estimate*** $\pm$ ***margin of error*** where the margin of error is a measure of the precision of the interval estimate (Larose, 2005).

Widely used estimation and prediction methods are:

- Point estimation,
- Confidence interval estimation,
- Linear and Multiple regression,
- Nonlinear regression,
- Logistic and Poisson regression.

Also decision trees, neural networks and k-Nearest Neighbour algorithms are used for estimation and prediction of the value of a sample. (Larose, 2005, Han & Kamber, 2001)

2.3.3.5 Association

Association rule mining searches for interesting relationships among the items in a data set. It is the study of attributes or characteristics that"go together". Association analysis is useful for discovering interesting patterns or relationships hidden in large data sets. The outcomes are represented in the form of *association rules* containing *if - then - else* statements. The strength of an association is measured in terms of its **support** and **confidence** (Han & Kamber, 2001, Larose, 2005). Support determines how often a rule is applicable to a given data set, and confidence shows how frequently items in Y appear in transactions that contain X. Simply formulating support and confidence; for an association like $A \Rightarrow B$ (if A then B)

$$support = P(A \cap B) = \frac{\textit{number of samples containing both A and B}}{\textit{total number of samples}} \tag{2.3.1}$$

and

$$confidence = P(B \mid A) = \frac{\textit{number of samples containing both A and B}}{\textit{number of samples containing A}} \tag{2.3.2}$$

Association rule mining is a two step process: (1) finding all frequent itemsets, (2) generating rules from the frequent itemsets. The first step determines the overall performance of the mining association rules.

Most widely used area of association rule mining is called market basket analysis. It investigates the shopping behaviours of the customer and provides new offers of products to them. Especially related items sold together, gives the market owner big advantages to display products under the title of "*You may also want to see...*" or "*People who bought this, also bought that...*".

Association analysis has a huge computational complexity. If there are k items in the data, there may be $k.2^{k-1}$ possible association rules. To overcome this complexity, number of frequent itemsets are reduced (*Apriori* principle), or the number of comparisons are reduced (Larose, 2005).

CHAPTER THREE
BIOMEDICAL SIGNAL SOURCES AND REAL LIFE PROBLEMS

3.1 Biomedical signals

Living organisms are made up of many component systems - the human body, for example include the nervous system, the cardiovascular system and the musculoskeletal system, among others. Each system is made up of several subsystems that carry on many *physiological processes.* For example, the cardiac system performs the important task of rhythmic pumping of blood throughout the body to facilitate the delivery of nutrients as well as pumping blood through the pulmonary system for oxygenation of the blood itself.

Physiological processes are complex phenomena, including nervous or hormonal stimulation and control; inputs and outputs that could be in the form of physical material, neurotransmitters, or information; and action that could be mechanical, electrical or biochemical. Most physiological processes are accompanied by or manifest themselves as *signals* that reflect their nature and activities. Such signals could be of many types, including biochemical in the form of hormones and neurotransmitters, electrical in the form of potential or current, and physical in the form of pressure or temperature (Rangayyan, 2002).

3.2 Biomedical Signal Samples

- **The action potential (AP)** is the electrical signal that accompanies the mechanical contraction of a single cell when stimulated by an electrical current and it is caused by the flow of Na^+, K^+, Cl^- and other ions across the cell membrane (Rangayyan, 2002).
- **The Electroneurogram (ENG)** is an electrical signal observed as a stimulus

and the associated nerve action potential propagate over the length of a nerve (Rangayyan, 2002).

- **The Electromyogram (EMG)** is a technique for evaluating and recording the activation signal of muscles. EMG is performed using an instrument called an electromyograph, to produce a record called an electromyogram. An electromyograph detects the electrical potential generated by muscle cells when these cells are mechanically active, and also when the cells are at rest (Wikipedia, 2009).

- **The Electrocardiogram (ECG)** is the electrical manifestation of the contractile activity of the heart, and can be recorded fairly easily with surface electrodes on the limbs or chest. The ECG is perhaps the most commonly known, recognized and used biomedical signal. The rhythm of the heart in terms of beats per minute (bpm) may be easily estimated by counting the readily identifiable waves.

- **The Electroencephalogram (EEG)** represents the electrical activity of the brain.

- **Event related potentials (ERPs)** includes the ENG and EEG in response to light, sound, electrical or other external stimuli.

- **The Electrogastrogram (EGG)**, the electrical activity of the stomach consists of rhythmic waves of depolarization and repolarization of its constituent smoothe muscle cells.

- **The Phonocardiogram (PCG)** is a vibrationor sound signal related to the contractile activity of the cardiohemic system (the heart and blood together).

- **The carotid pulse (CP)** is a pressure signal recorded over the carotid artery as it passes near the surface of the body at the neck.

- **Signals from catheter-tip sensors**: For very specific and close monitoring of the cardiac function, sensors placed on catheter tips may be inserted into the cardiac chambers. It then becomes possible to acquire several signals such as left ventricular pressure, right atrial pressure, aortic pressure and intracardiac sounds. While these signal provide valuable and accurate information, the procedures are invasive and are associated with certain risks.

- **The speech signal** is an important signal although it is more commonly considered as a communication signal than a biomedical signal. However, the speech signal can serve as a diagnostic signal when speech and vocal-tract disorders need to be investigated.
- **The vibromyogram (VMG)** is the direct mechanical manifestation of contraction of a skeletal muscle and is a vibration signal that accompanies the EMG.
- **The vibroarthogram (VAG)** is the vibration signal recorded from a joint during movement of the joint. Detection of knee-joint problems via the analysis of VAG signals could help avoid unnecessary exploratory surgery and also aid better selection of patients who would benefit from the surgery.
- **Oto-acoustic emission signals** represent the acoustic energy emitted by the cochlea either spontaneously or in response to an acoustic stimuli.

3.3 Objectives of Biomedical Signals

The representation of biomedical signals in electronic form facilitates computer processing and analysis of the data. Figure 3.1 illustrates the typical steps and processes involved in computer-aided diagnosis and therapy based upon biomedical signal analysis. The major objectives of biomedical instrumentation and signal analysis introduce in Rangayyan (2002) are:

- Information gathering - measurement of phenomena to interpret a system.
- Diagnosis - detection of malfunction, pathology or abnormality.

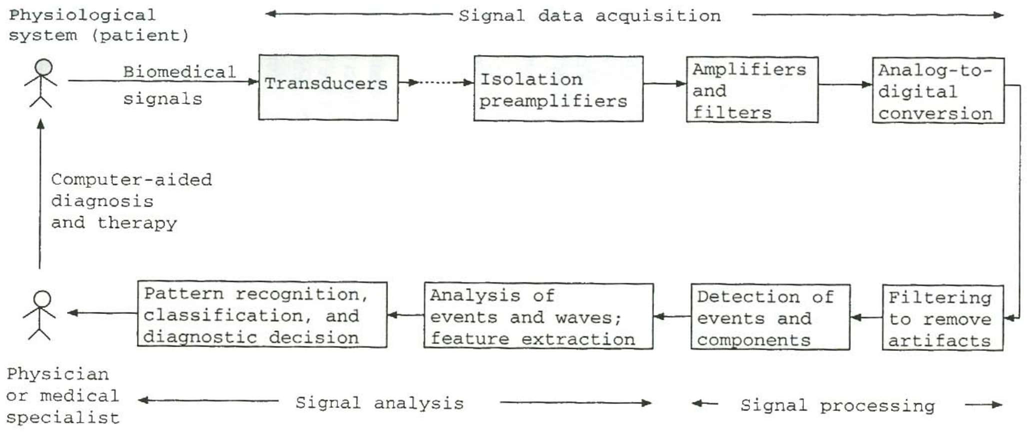

Physiological system (patient)
Signal data acquisition
Biomedical signals
Transducers
Isolation preamplifiers
Amplifiers and filters
Analog-to-digital conversion
Computer-aided diagnosis and therapy
Pattern recognition, classification, and diagnostic decision
Analysis of events and waves; feature extraction
Detection of events and components
Filtering to remove artifacts
Physician or medical specialist
Signal analysis
Signal processing

Figure 3.1 Computer aided diagnosis and therapy based upon biomedical signal analysis (Rangayyan, 2002)

- Monitoring - obtaining continuous or periodic information about a system.
- Therapy and control - modification of the behavior of a system based upon the outcome of the activities listed above to ensure a specific result.
- Evaluation - objective analysis to determine the ability to meet functional requirements, obtain proof of performance, perform quality control or quantify the effect of treatment.

3.4 Difficulties in Biomedical Signals

In spite of the long history of biomedical instrumentation and its extensive use in health care and research, many practical difficulties are encountered in biomedical signal acquisition, processing and analysis. The characteristics of the problem and hence their potential solutions are unique to each type of signal. Particular attention should be paid to the following issues according to Rangayyan (2002):

- Accessibility of the variables to measurement.
- Variability of the signal source.

- Inter-relationship and interactions among physiological systems.
- Effect of the instrumentation or procedure on the system.
- Physiological artifacts and interference.
- Energy limitations.
- Patient safety.

3.5 Brain and EEG

Human brain is one of the most critical organs for the human body. It is located in the most secured region with a closed cap of bones (skull) of the body. It is named *encephalon* in Latin which comes from ancient Greek word *enkephalos* - in the head. It is the center of learning and it regulates thought, memory, judgement, personal identity, and other aspects of what is commonly called the mind. It also regulates aspects of the body - including body temperature, blood pressure and the activity of internal organs - to help the body respond to its environment and to remain healthy. The brain is said to be the most complex living structure known to the universe (Britannica, 2008).

The brain and the spinal cord make up the central nervous system processing and communicating the information that controls all of the body functions. The spinal cord extends from the base of the brain and is contained within the vertebral canal. The brain controls the activities of the body and receives information about the body's inner workings, and about the outside world by sending and receiving signal via the spinal cord and the peripheral nervous system. It receives the oxygen and foot it needs to function by way of a vast network of arteries that carries fresh blood to every part of the brain.

The brain of a human adult weights about 1 - 1.5 kg. with a volume of 1600 cm^3. It consumes 20% - 25% of the overall energy produced by the body. In a typical human the cerebral cortex (the largest part) is estimated to contain 15 to 33 billion neurons, each connected by synapses to several thousand other neurons. These neurons communicate with one another by means of long protoplasmic fibers called axons, which carry trains of signal pulses called action potentials to distant parts of the brain or body targeting specific recipient cells (Wikipedia, 2012).

The brain looks like a mushroom contained within the skull.The cap of the mushroom is the *cerebrum* and the stem of the mushroom (the part attached to the spinal cord) is the *brainstem.* At the back of the head between the brainstem and the cerebrum is the *cerebellum.*

The ***cerebrum*** is the largest and most highly developed part of the brain. It is divided into four sections or lobes:

- **Frontal lobe** controls cognitive functions such as speech, planning and problem solving,
- **Parietal lobe** is assigned for controlling sensation such as touch, pressure and judging size and shape,
- **Temporal lobe** mediates visual and verbal memory, and smell,
- **Occipital lobe** controls visual reception and recognition of shapes and colors.

Symmetrical in structure, the cerebrum is divided into the left and right hemispheres. In most people, the left hemisphere is responsible for functions such as creativity, and the right hemisphere is responsible for functions including logic and spatial perception. The left hemisphere controls the movement of the right half of the body, and the right hemisphere controls the movement of the left half of the body. This is because the nerve fibres that send messages to the body cross over in the medulla, part of the brainstem (Britannica, 2008).

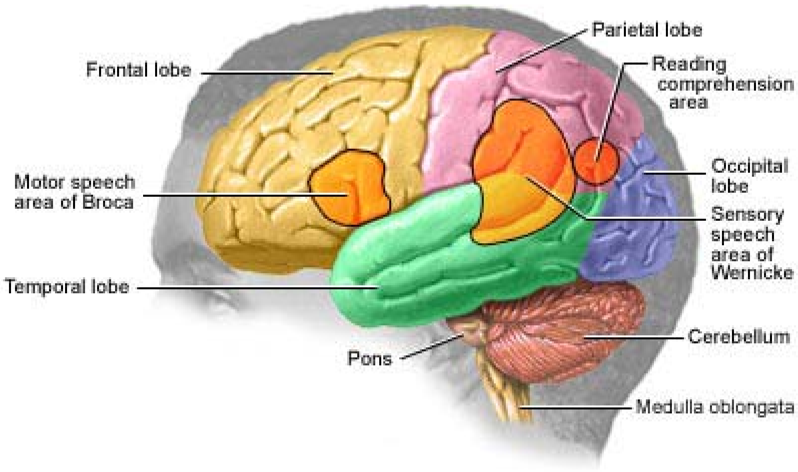

Frontal lobe
Parietal lobe
Reading comprehension area
Motor speech area of Broca
Occipital lobe
Sensory speech area of Wernicke
Temporal lobe
Cerebellum
Pons
Medulla oblongata

Figure 3.2 Basic parts of human brain

The most prominent series of observations clearly belonging to modern neuropsychology was made by Paul Broca in the 1860s. He reported the cases of several patients whose speech had been affected following damage to the left frontal lobe and provided autopsy evidence of the location of the lesion. Broca explicitly recognized the left hemisphere's control of language, one of the fundamental phenomena of higher cortical function.

In 1874 the German neurologist Carl Wernicke described a case in which a lesion in a different part of the left hemisphere, the posterior temporal region, affected language in a different way. In contrast to Broca's cases, language comprehension was more affected than language output. This meant that two different aspects of higher cortical function had been found to be localized in different parts of the brain. In the next few decades there was a rapid expansion in the number of cognitive processes studied and tentatively localized.

Wernicke was one of the first to recognize the importance of the interaction between connected brain areas and to view higher cortical function as the build-up of complex mental processes through the coordinated activities of local regions dealing with relatively simple, predominantly sensory-motor functions. In doing so, he opposed the view of the brain as an equipotential organ acting en masse.

Broca's declaration that the left hemisphere is predominantly responsible for language-related behaviour is the clearest and most dramatic example of an asymmetry of function in the human brain. This functional asymmetry is related to hand preference and probably to anatomical differences, although neither relationship is simple (Britannica, 2008).

3.6 Brain Data Measurements

Human brain has always been an attractive body part for researchers because of its functional complexity and large function spectrum. Many different techniques have been developed for detecting anomalies or damages as well as understanding how brain works. Some of these techniques as invasive. High levels of anatomical and metabolic data can be provided with different brain imaging techniques. These techniques are as follows:

Electroencephalogram (EEG) techniques date back to the work of Canton with animals in the 1800's and that of Berger with humans in the 1920's. The basic idea is to use activity recorded from the scalp as a window to underlying brain processing. Technically, EEG measures the difference in the brain's electrical activity found between two electrodes. EEG will be mentioned in detail in the next section.

Event-related potentials (ERPs), as the name implies, show EEG activity in relation to a particular event. ERPs have been used to reflect the processing of cognitive, emotional, and sensory stimuli in the brain. EEG and ERPs have a real value in determining the time course of a response, because they reflect millisecond changes within the electrical activity of the cortex (Ray & Oathes, 2003).

The ***MagnetoEncephaloGram (MEG)*** uses SQUID (Superconducting Quantum Interference Device) to detect the small magnetic field gradients exiting and entering the surface of the head that are produced when neurons are active. MEG signals are similar to EEG signals but have one important advantage: magnetic fields are not distorted when they pass through the cortex and the skull, which makes localization of sources more accurate than EEG (Ray & Oathes, 2003).

Computerized Axial Tomography (CAT), or computerized tomographic imaging is a diagnostic imaging method using a low-dose beam of X-rays that crosses the body in a single plane at many different angles. A major advance in imaging technology, it became generally available in the early 1970s. The technique uses a tiny X-ray beam that traverses the body in an axial plane. Detectors record the strength of the exiting X-rays, and that information is then processed by computer to produce a detailed two-dimensional cross-sectional image of the body. A series of such images in parallel planes or around an axis can show the location of abnormalities and other space-occupying lesions (especially tumours and other masses) more precisely than can conventional X-ray images (Encyclopaedia Britannica, 2012).

Positron emission tomography (PET) systems measure variations in cerebral blood flow that are correlated with brain activity. It is through blood flow that the brain obtains oxygen and glucose from which it gets its energy. By measuring changes in blood flow in different brain areas, it is possible to infer which areas of the brain are more or less active during particular tasks (Ray & Oathes, 2003).

Like PET, ***functional Magnetic Resonance Imaging (fMRI)*** is based on the fact that blood flow increases in active areas of the cortex. However, it uses a different technology from PET in that in fMRI local magnetic fields are measured in relation to an external magnet. Specifically, hemoglobin, which carries oxygen in the bloodstream, has different magnetic properties before and after oxygen is absorbed. Thus, by measuring the ratio of hemoglobin with and without oxygen, the fMRI is able to map changes in cortical blood and infer neuronal activity (Ray & Oathes, 2003).

Near InfraRed Spectroscopy (NIRS) is an optical technique for measuring blood oxygenation in the brain. It works by shining light in the near infrared part of the spectrum (700-900nm) through the skull and detecting how much the remerging light is attenuated. How much the light is attenuated depends on blood oxygenation and thus NIRS can provide an indirect measure of brain activity (Demitri, 2007).

3.7 ElectroEncephaloGraphy - EEG

An early discovery established that the brain is associated with the generation of electrical activity. Richard Caton had demonstrated already in 1875 that electrical signals in the microvolt range can be recorded on the cerebral cortex of rabbits and dogs. Several years later, Hans Berger recorded for the first time electrical "brain waves" by attaching electrodes to the human scalp; these waves displayed a time-varying, oscillating behaviour that differed in shape from location to location on the scalp. Berger made the interesting observation that brain waves differed not only between healthy subjects and subjects with certain neurological pathologies, but that the waves were equally dependent on the general mental state of the subject, e.g., whether the subject was in a state of attention, relaxation, or sleep. The experiments conducted by Berger became the foundation of *electroencephalography*, later to become an important noninvasive clinical tool in better understanding the human brain and for diagnosing various functional brain disturbances (Sörnmo & Laguna, 2005).

Electroencephalography (EEG) is a graphical display of a difference in voltages from two sites of brain function recorded over time. Electroencephalography involves the study of recording these electrical signals that are generated by the brain via a cap with electrodes. Most routine EEGs recorded at the surface of the scalp represent pooled electrical activity generated by large numbers of neurons. Electrical signals are created when electrical charges move within the central nervous system. Neural function is normally maintained by *ionic gradients* established by neuronal membranes. Sufficient duration and length of small amounts (in microvolts) of electrical currents of cerebral activity are required to be amplified and displayed for interpretation (Tatum et al., 2007).

Signals recorded from the scalp have, in general, amplitudes ranging from a few microvolts to approximately 100 μV and a frequency content ranging from 0.5 to 30-40 Hz. Electroencephalographic signal frequencies are conventionally classified into five different frequency bands: Delta (0.5 - 4 Hz.), Theta (4-7 Hz.), Alpha (8-14 Hz.), Beta (15-30 Hz.) and Gamma (>28 Hz.) (Sörnmo & Laguna, 2005, Megalooikonomou et al., 2000, Tatum et al., 2007, Bayazıt, 2009, Öniz, 2006).

EEG data can be used for many purposes. *Spontaneous activity* is measured on the scalp or on the brain and is called the electroencephalogram. The amplitude of the EEG is about 100 μV when measured on the scalp, and about 1-2 mV when measured on the surface of the brain. The bandwidth of this signal is from under 1 Hz to about 50 Hz. As the phrase “spontaneous activity” implies, this activity goes on continuously in the living individual. *Evoked potentials* are those components of the EEG that arise in response to a stimulus (which may be electric, auditory, visual, etc.) Such signals are usually below the noise level and thus not readily distinguished, and one must use a train of stimuli and signal averaging to improve the signal-to-noise ratio. *Single-neuron behaviour* can be examined through the use of microelectrodes which impale the cells of interest. Through studies of the single cell, one hopes to build models of cell networks that will reflect actual tissue properties (Malmivuo & Plonsey, 1995).

3.7.1 Recording EEG

EEG recordings are received via a cap worn on the head. There are conductive receivers called *electrode* on the cap touching the surface of the skull. Mostly an inductive gel is injected in each electrode to increase the sensitivity. Each electrode is connected to an amplifier by wires. The amplifier makes the signal stronger and transmits the signal info to a computer or a scrolling paper. The clinical EEG is commonly recorded using the International 10/20 system, which is a standardized system for electrode placement. This particular recording system (electrode montage) employs 21 electrodes attached to the surface of the scalp at locations defined by certain anatomical reference points; the numbers 10 and 20 are percentages signifying relative distances between different electrode locations on the skull perimeter (see Figure 3.3 and 3.4 presented in Malmivuo & Plonsey (1995)). Note that odd-numbered electrodes are on the left side and even-numbered electrodes are on the right side. Z (zero) is the the midline (Sörnmo & Laguna, 2005).

3.7.2 EEG Applications

EEG is a non-invasive, simple (in proportion to other techniques) and instant method for brain data capturing. Many applications and researches depend on studies in EEG analysis. Investigating EEG signals, some disorders can be diagnosed, especially in epilepsy and sleep disorders - which is the two of the most important clinical applications of EEG analysis. .

Epilepsy is caused by several pathological conditions such as brain injury, stroke, brain tumours, infections, and genetic factors. The EEG is the principal test for diagnosing epilepsy and gathering information about the type and location of seizures (Sörnmo & Laguna, 2005).

Sleep disorders, which are frequent in our society, may be caused by several conditions of medical and/or psychological origin. There are 4 groups of sleep disorders defined in Sörnmo & Laguna (2005): insomnia, hypersomnia, circadian rhythm disorders, parasomnia. EEG is one of the favourite methods used in sleep disorder studies.

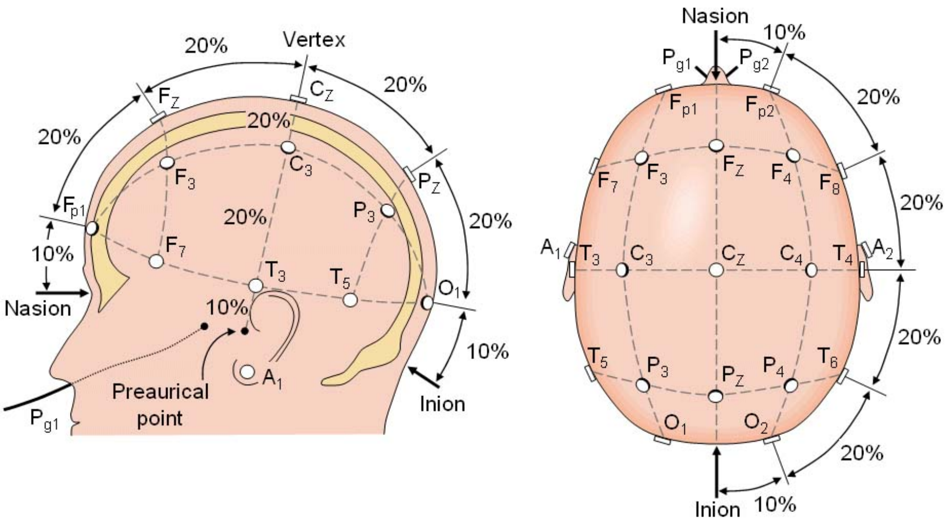


Figure 3.3 Electrode locations for international 10-20 system

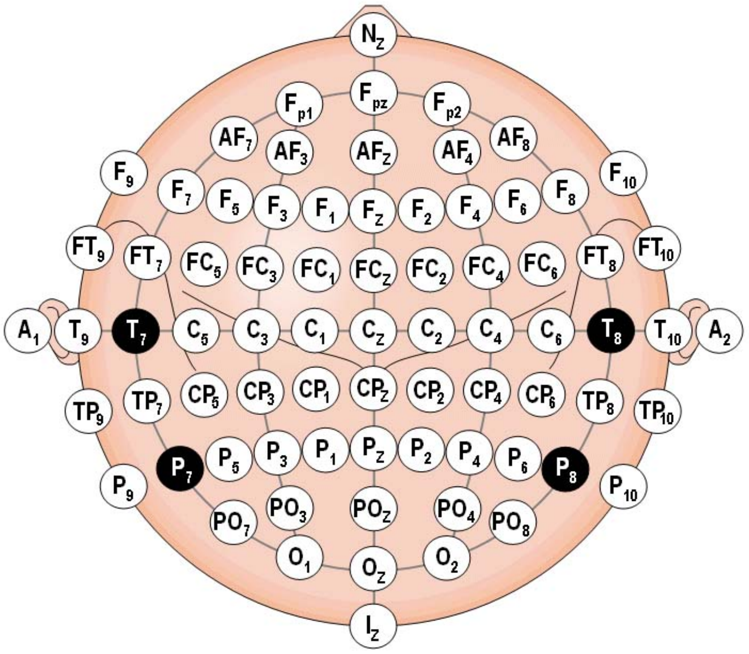


Figure 3.4 A = Ear lobe, C = central, Pg = nasopharyngeal, P = parietal,
F = frontal, Fp = frontal polar, O = occipital.

EEG is also used to help for diagnosing brain seizures/diseases and their type. These include abnormal changes in body chemistry that affect the brain, brain diseases such as Alzheimer, infections or tumours in the brain. Additionally, EEG is used to monitor the depth of anesthesia, and to detect the brain death.

EEG and ERPs are used in neuroscience, cognitive science and psychology, biophysics and psychophysiological researches (Wikipedia, 2006). EEG signals are helpful for detecting structural and functional asymmetry of the brain and mapping or localization studies.

3.8 Dichotic Listening

Dichotic listening has been used in hundreds of research and clinical reports related to language processing, emotional arousal, hypnosis and altered states of consciousness, stroke patients, psychiatric disorders and child disorders, including dyslexia and congenital hemiplegia. One frequently used method to study language asymmetry is dichotic listening. Because of its ability to distinguish which hemisphere processes specific sounds, the use of dichotic listening has become widespread in studies of brain asymmetry (Hugdahl, 2005).

Dichotic listening is applied by presenting two auditory stimuli simultaneously, one in each ear, through earphones. The subject reports which of the two stimuli was perceived best. The test follows a typical sequence of events, in which a dichotic stimuli is presented followed by the subject reporting what he heard, usually out of a list of six syllables (*ba, da, ga, pa, ta, ka*) or two tones. The signals presented to the subject to the left ear (LE) and right ear (RE) are compared with the response of the subject. Most common approaches for the outcomes is counting or calculating percentage values of the true responses. The difference of RE and LE describes the ear advantage of the subject (REA, LEA or NoEA) (Kent, 2003).

CHAPTER FOUR
APPLICATION

4.1 Data Mining and EEG

The problem of multidimensional data (e.g. brain images), can be solved with newer mining methods which are applied directly to the images in order to capture most of their information content. Data mining is heavily dependent on statistical methods for discovering associations and classifications among disparate types of data. EEG technique seems wealthier to examine from data mining perspective because of the following advantages:

- EEG data has a high time resolution configured by the recorder. Different sampling rates can be applied. As other methods for researching brain activity have time resolution between seconds and minutes, the EEG has a resolution down to sub-milliseconds.
- Electric activity is easy to measure. By using a number of electrodes and different numbered caps, the electrical potential differences can be measured spontaneously from the head without any intervention to the subject.
- Recording EEG does not rely on blood flow or metabolism. Other methods for exploring functions in the brain require blood flow or metabolism. Newer research typically combines EEG or MEG with MRI or PET to get high temporal and spatial resolution.
- EEG provides spontaneous measurement of a response of a subject for a specific interaction (like stimuli) or event (like an epileptic attack). To see the result of a stimulus, no need to wait for the result of analysis like blood test or something similar.
- EEG data can be combined with other body function measures. To measure

the correct respond to a stimulus, EEG records can be analysed parallel with some other body measures like heart bit rate, blood pressure, etc.

- EEG comes in large databases suitable for data mining operations. For example, one whole night recording of the human sleep results in 8 h of multichannel data sampled with up to 256 Hz.

There are also some disadvantages working with EEG data. These can be listed as below(Flexer, 2000, Megalooikonomou et al., 2000, Sörnmo & Laguna, 2005):

- EEG signals are very noisy. Whereas the electrical background activity of the human brain is in the range of 1 - 200 μV, evoked potentials (EPs) have amplitude of only 1 - 30 μV.
- EEG signals have a large temporal variance. Although the spatial localization of EEG is already well researched, a lot of effort is still needed to take the between-subjects temporal variation into account.
- Analysis of EEG data requires the use of the full range of data mining techniques besides the signal processing operations. The signals must cleaned, be transformed into different domains (frequency, time) and must be filtered. There are tasks for classification, regression, clustering, sequence analysis, etc. for investigating EEG data.

4.2 The Experiment - Business Understanding

In the application section of the thesis, EEG data were obtained from Dokuz Eylül University, Department of Biophysics, Research Laboratory for PhD thesis of Onur Bayazıt (Bayazıt, 2009). The data perceived from different subjects contain the EEG recordings under the dichotic listening test.

A total of 60 healthy subjects (behavioral main group; mean age 23.38 years, 30 female) participated voluntarily in the DL study after having given informed written consent. A subgroup of 20 subjects (mean age: 21.15, 10 females) formed the electrophysiological subject pool. The subjects were mainly students at the University of Dokuz Eylül University Medical Faculty, İzmir. The subjects reported no history of any neurological and psychiatric conditions and all were native Turkish speakers (Bayazıt et al., 2009).

Data used in this study was obtained in the specific experiment made by Dokuz Eylül University Department of Brain Biophysics laboratories and contains the unfiltered EEG recordings of a subject captured by 64 electrodes cap.

In the experiment, subject is given a dichotic stimulus (combination of two consonant vowel syllables like *BA* to the left ear and *DA* to the right ear) at pseudo-random time. 2170 msec later than the stimulus, a light indicator is lit to inform the subject to answer about what was heard. The answer keypad contains 6 buttons each assigned to declare vowel syllables *ba, da, ga, ka, pa, ta*. The subject presses the related button and again a pseudo-random time passes, second stimuli is delivered. 36 different pairs of stimuli are applied twice to the subject. As a result there are 60 stimulus with two different syllables and 12 homonym stimulus containing the same syllables. During this procedure, EEG recordings are received from 64 electrodes of the subject. Continuous EEG activity was taken with a sampling rate of 1 kHz, filtered between 0.15 and 70 Hz (Bayazıt et al., 2009, Vahaplar et al., 2011).

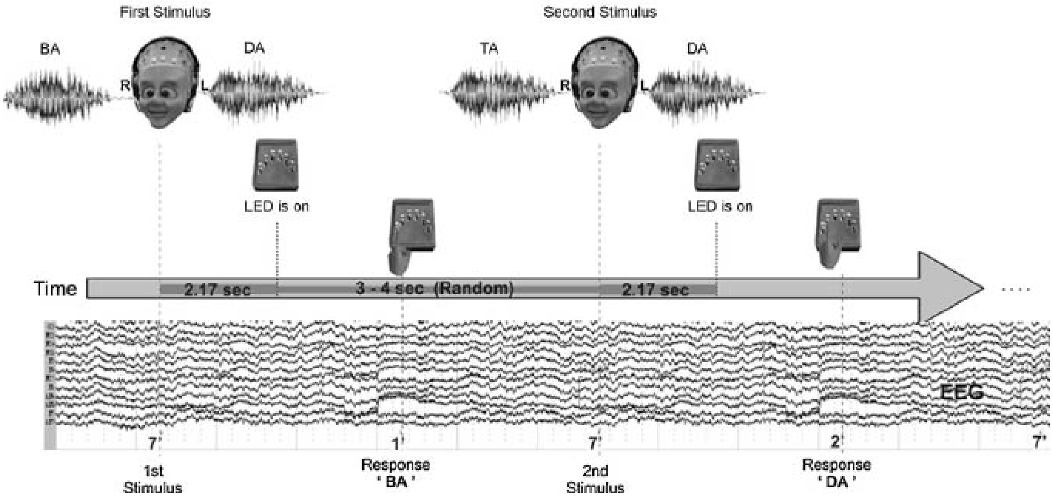


Figure 4.1 EEG recording during dichotic listening task, captured from (Bayazıt et al., 2009)

4.3 Summarizing Data - Data Understanding

Mainly 4 data sets were used in this thesis study. 2 of the sets (Data Set 1 and Data Set 3) have Left Ear Advantage (LEA) and the other two (Data Sets 2 and Data Set 4) have Right Ear Advantage (REA). The data sets were obtained in MATLAB .mat file format containing the following information:

- **Data:** 64 x (size of record in milliseconds). (Rows indicate the electrode number, columns are the voltage measurements received in that milliseconds. Data(15,123456) stores the voltage value measured in 123456th millisecond on the 15th electrode).
- **Event:** Information about the events (stimuli and response) labeled on EEG data.
  - *Type:* type of the event labelled on the data. 7 means the stimulus, 1,2,3,4,5,6 are the responses corresponding to the syllables *ba, da, ga, ka, pa, ta* respectively.
  - *Latency:* It is the time information of the event when occured (in milliseconds).

  - *Urevent:* Order of the event. (Totally 144 - 72 for stimuli, 72 for response)

- **Chanlocs:** Electrode labels of the cap (Ex: 15 denotes ‘CZ’ electrode).

4.4 Preliminary Work - Data Preprocessing

As the first step of the study, some functions to partition the EEG data into individual pieces were written in MATLAB. This code splits the whole data into the stimuli-based pieces. EEG recordings for each stimuli were separated and a new matrix was constructed for each stimuli for the selected electrode. Then each stimuli data were divided into three sections: *pre_stimuli, post_early and post_late*. pre_stimuli section is the part of the recordings beginning from the last response until the stimuli is given. post_early section contains the recordings and starts from the time that the stimuli is given and ends until the light indicator comes. post_late section is the part beginning at the time that the light indicator came up to the subject presses the button.

Pre_stimuli and post_late sections have variable lengths in time due to the response of the subject for each stimuli but post_early section has a fixed size of 2170 milliseconds (stimuli and light indicator time interval). The main target of the study was on the post_early section where the stimuli effects are observed in the brain.

Initially, the response times of stimuli answers were analysed for each of the four datasets. The stimuli Negative response times mean that subject has pressed the button on the keypad before the light indicator was lit. Wrong answers are the ones which subject responded but the stimuli does not contain that syllable answered (ex: the stimulus was BA-DA, but the subject responded KA). Table 4.1 gives the descriptives of the data sets.

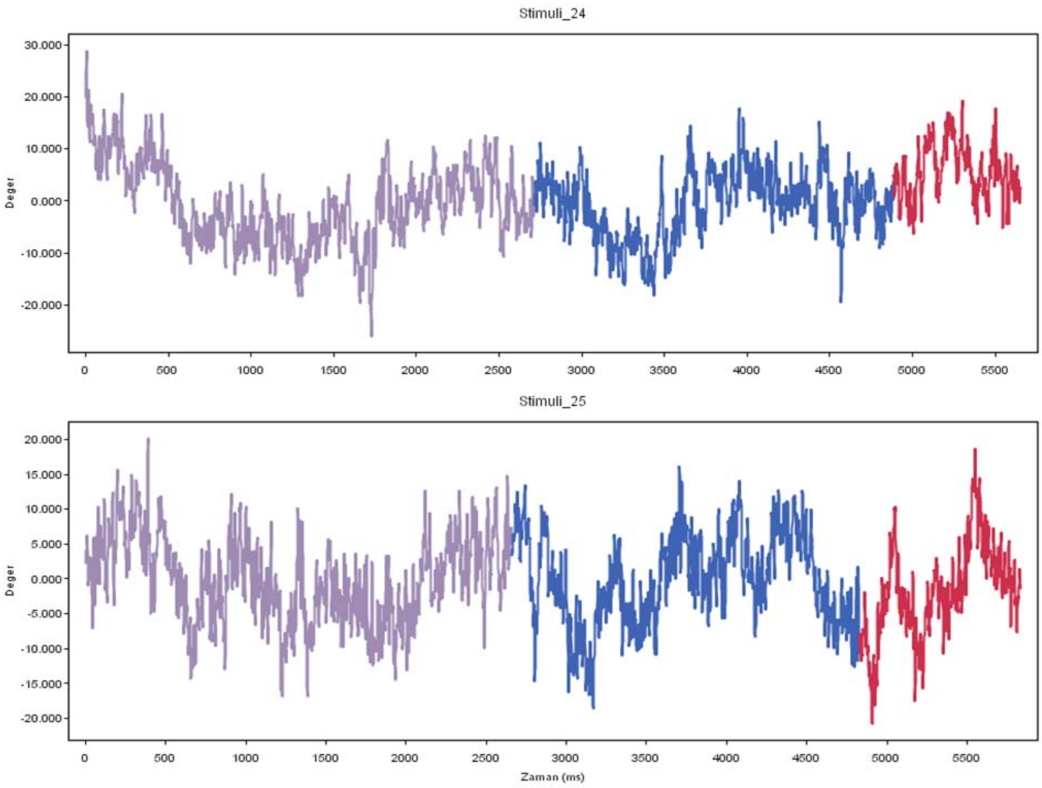


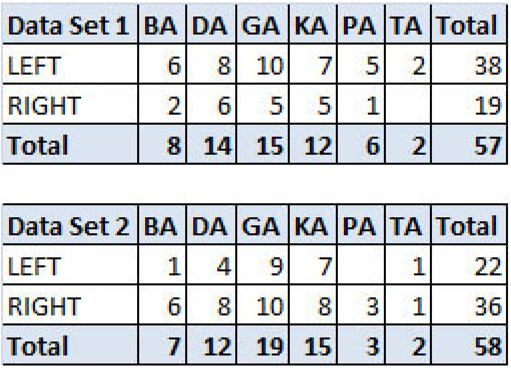
Figure 4.2 Labeled EEG recordings of two stimuli

Table 4.1 Average, minimum and maximum response times of four subjects

Data Sets	Min	Max	Avg	Left	Right	Wrong	L%	R%	W%
Data Set 1	-125	2123	719.14	38	19	3	63%	32%	5%
Data Set 2	-142	1267	561.06	22	36	2	37%	60%	3%
Data Set 3	822	2354	1403.80	30	21	9	50%	35%	15%
Data Set 4	-439	1169	420.38	10	43	7	17%	72%	12%

As an example, for Data Set 1, the subject has responded the stimuli number 3, 8, 9, 10, 12, 13, 14, 15, 18, 19, 20, 21, 24, 25,26, 30, 31, 32, 35, 39, 44, 45, 46, 48, 49, 50, 51, 54, 55, 56, 57, 60, 61, 64, 66, 67, 68, 71 with Left Ear and stimuli 1, 2, 4, 7, 16, 22, 27, 28, 33, 36, 37, 38, 40, 43, 52, 58, 62, 63, 72 with Right Ear. The stimuli 5, 11, 17, 23, 29, 34, 41, 47, 53, 59, 65, 70 are the homonym ones (same syllable on left and right). Subject 1 has wrong responses on stimuli 6, 42 and 69.

At this stage, the responded syllables were examined in terms of response counts. In Figure 4.3 and Figure 4.4, response counts by ear sides are given. It can be seen that Data Set 1 has no responses of *'TA'* in right ear stimuli and Data Set 2 has no responses of *'PA'* in left ear stimuli.



Data Set 1	BA	DA	GA	KA	PA	TA	Total
LEFT	6	8	10	7	5	2	38
RIGHT	2	6	5	5	1		19
Total	**8**	**14**	**15**	**12**	**6**	**2**	**57**

Data Set 2	BA	DA	GA	KA	PA	TA	Total
LEFT	1	4	9	7		1	22
RIGHT	6	8	10	8	3	1	36
Total	**7**	**12**	**19**	**15**	**3**	**2**	**58**

Figure 4.3 Syllables and Responses on Data Set 1 and 2

Then, the signals were grouped by the response syllables and signal averages of 0-300 msec. were calculated. Group averages are given in Figure 4.5.

Next, the signals that the subject responded by left and right ear were grouped. The signals responded with right ear and left ear, wrong responses and homonym stimuli were grouped and signal averages were computed. The resulting table is given in the following (Figure 4.6).

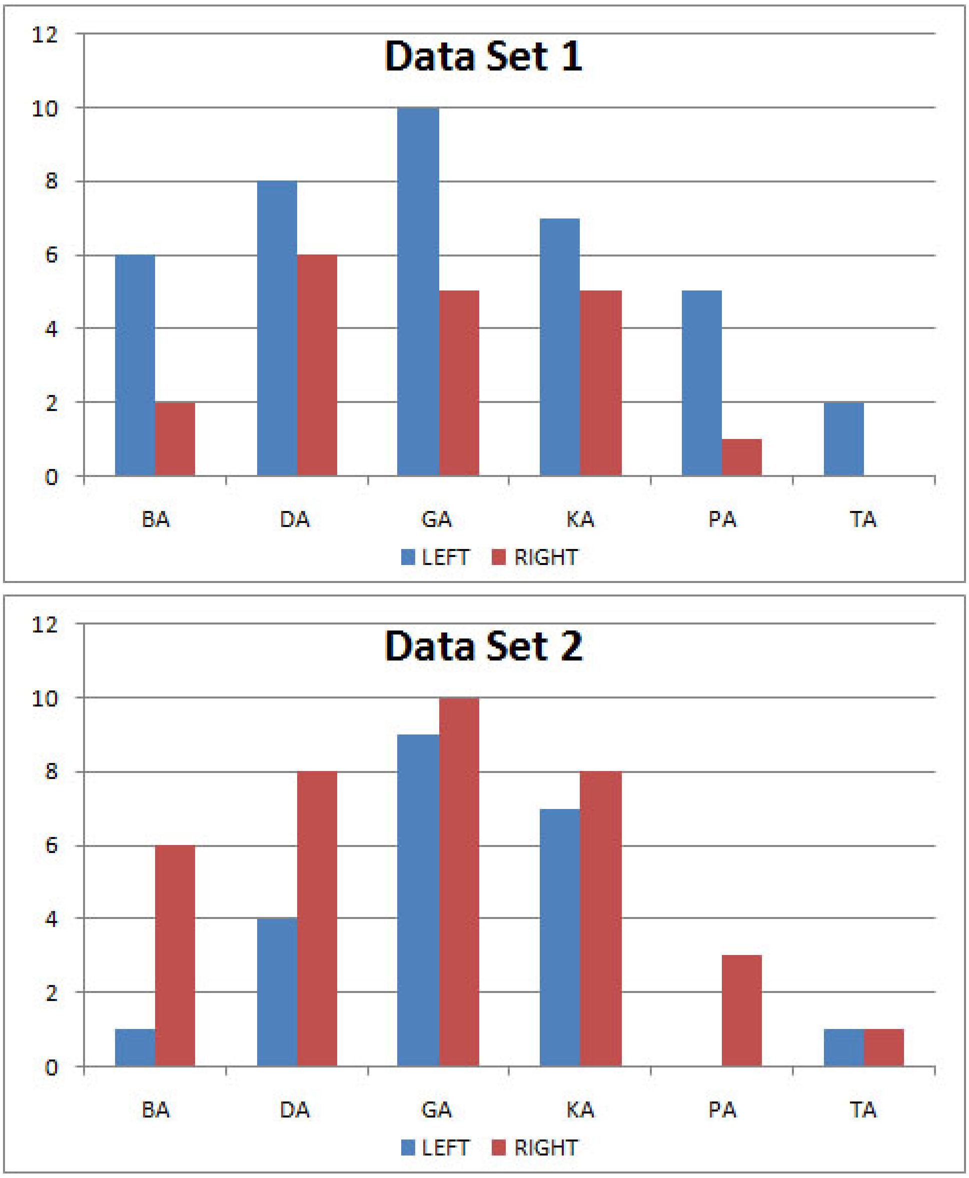
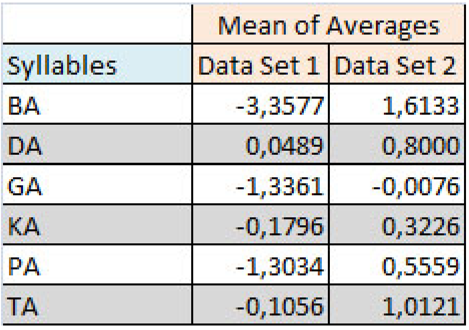


Figure 4.4 Bar graph of syllables and responses on Data Set 1 and 2

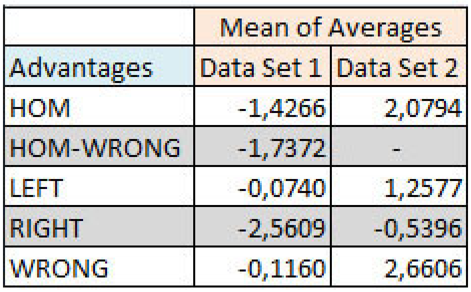
As another grouping, the syllables were labelled as “HARD” for *ka, pa, ta* and “SOFT” for *ba, da, ga*. The subject’s responses were examined according to the type of the response. Figure 4.7 displays the response counts of the subjects in HARD and SOFT types.

To detail this grouping, the distributions of HARD and SOFT responses sent to the ears were analysed. Left and Right syllables were labelled as HARD and SOFT,



	Mean of Averages	
Syllables	Data Set 1	Data Set 2
BA	-3,3577	1,6133
DA	0,0489	0,8000
GA	-1,3361	-0,0076
KA	-0,1796	0,3226
PA	-1,3034	0,5559
TA	-0,1056	1,0121

Figure 4.5 Signal averages of syllable responses on Data Set 1 and 2



	Mean of Averages	
Advantages	Data Set 1	Data Set 2
HOM	-1,4266	2,0794
HOM-WRONG	-1,7372	-
LEFT	-0,0740	1,2577
RIGHT	-2,5609	-0,5396
WRONG	-0,1160	2,6606

Figure 4.6 Signal averages of stimuli based on responding ear

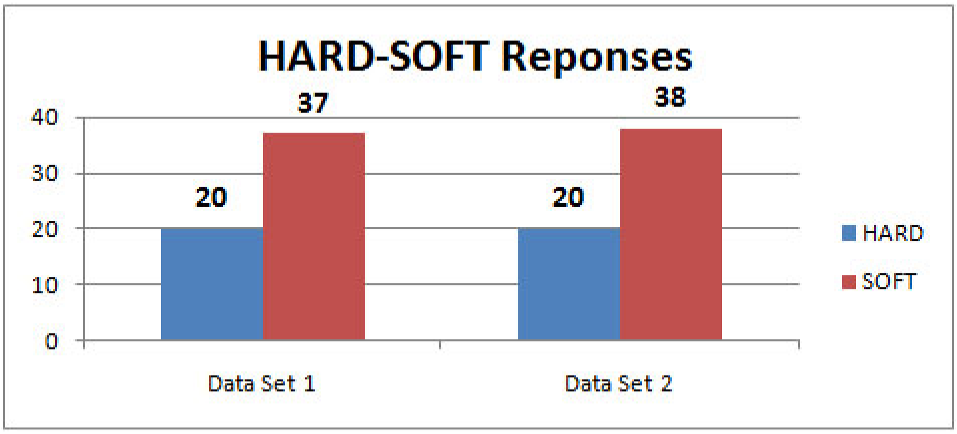


Figure 4.7 Response counts of subjects in HARD and SOFT syllables

and the responses were compared according to the placement of the syllable type given to each ear. Also the ear distributions are shown in Figures 4.8 and 4.9. The *Stim Type* denotes the type of the syllable in the stimuli for the corresponding ear (ex: HARD-SOFT means a HARD syllable was presented in the Left ear and a SOFT one in the right ear).

Data Set 1	Responses		Data Set 2	Responses	
Stim Type	**HARD**	**SOFT**	**Stim Type**	**HARD**	**SOFT**
HARD-HARD	18		HARD-HARD	18	
HARD-SOFT	9	9	HARD-SOFT	3	15
SOFT-HARD	2	16	SOFT-HARD	7	11
SOFT-SOFT		18	SOFT-SOFT		18
TOTAL	**29**	**43**	**TOTAL**	**28**	**44**

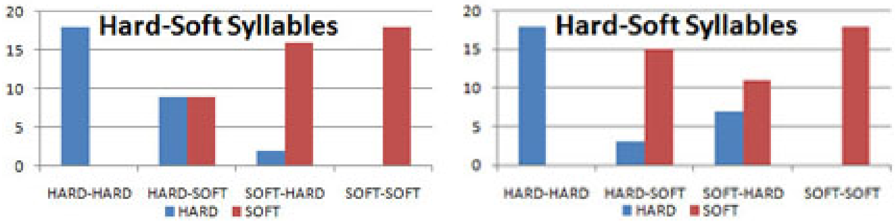


Figure 4.8 Response types of subjects in HARD and SOFT syllables

Data Set 1	Responses			Data Set 2	Responses		
Stim Type	**LEFT**	**RIGHT**	**TOTAL**	**Stim Type**	**LEFT**	**RIGHT**	**TOTAL**
HARD-HARD	7	4	11	HARD-HARD	5	6	11
HARD-SOFT	7	9	16	HARD-SOFT	3	15	18
SOFT-HARD	16	2	18	SOFT-HARD	11	6	17
SOFT-SOFT	8	4	12	SOFT-SOFT	3	9	12
TOTAL	**38**	**19**	**57**	**TOTAL**	**22**	**36**	**58**

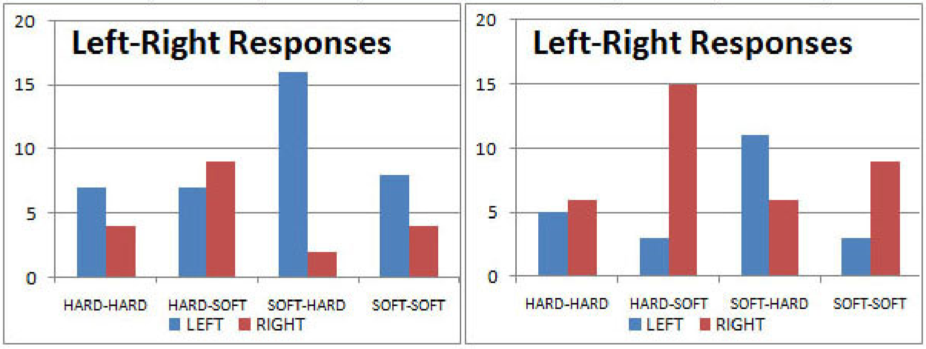


Figure 4.9 Distribution of HARD and SOFT syllables

4.5 Similarity Analysis

The section of the signal subject to this part of the study is the first 300 msecs. of Post_Early section. The recordings showed that the first effect of the stimuli appears in this time slice. The aim is to investigate the similarity or dissimilarity of the Right Ear and Left Ear responses and similarity of different electrodes.

Before working on the data, 0-300 msec. sections of each stimuli was selected. These were grouped according to the ear advantage of the subject's responses. The signals of stimuli which the subject responded with his/her right ear (REA) and left ear (LEA) were summed and averaged. Homonym stimuli (HOM) were also grouped for comparisons. The same procedure was applied for all electrodes. Figure 4.10 shows the average values of EEG recordings for the first subject (Data Set 1).

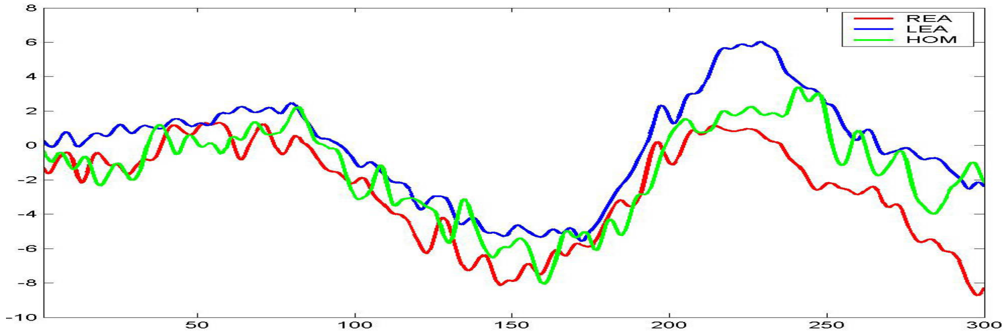


Figure 4.10 REA, LEA and HOM Averages of Data Set 1

Electrodes *CZ, C3, C4, F3, F4, T7* and *T8* were selected to compare. The reason for choosing these electrodes is that they are located on the hearing and linguistic processing parts of the brain as mentioned in section 3.5. The location of the electrodes are given in Figure 4.11.

For the similarity measure, the method of Z_M mentioned in section 2.2.2 (and introduced in Kennedy (2007)) was applied for the signals on each electrodes in doubles. Two different electrodes were put in signal similarity process. Each signal was formed of a 1x300 vector of recorded EEG measurement corresponding to the

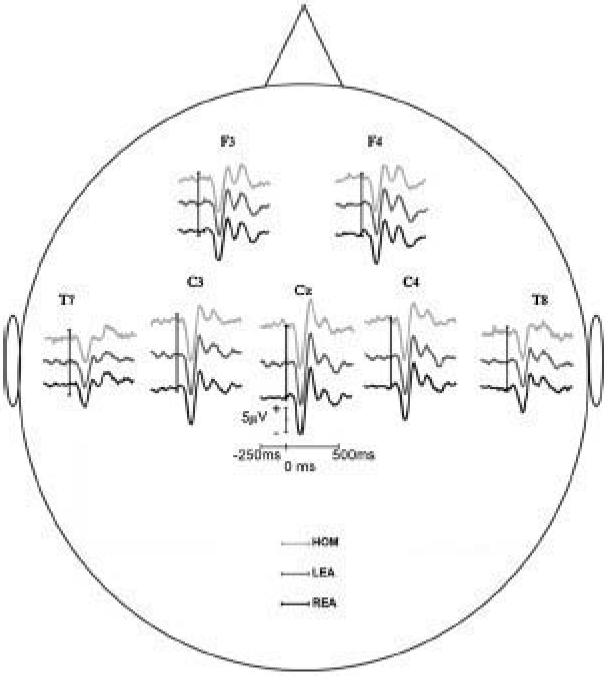
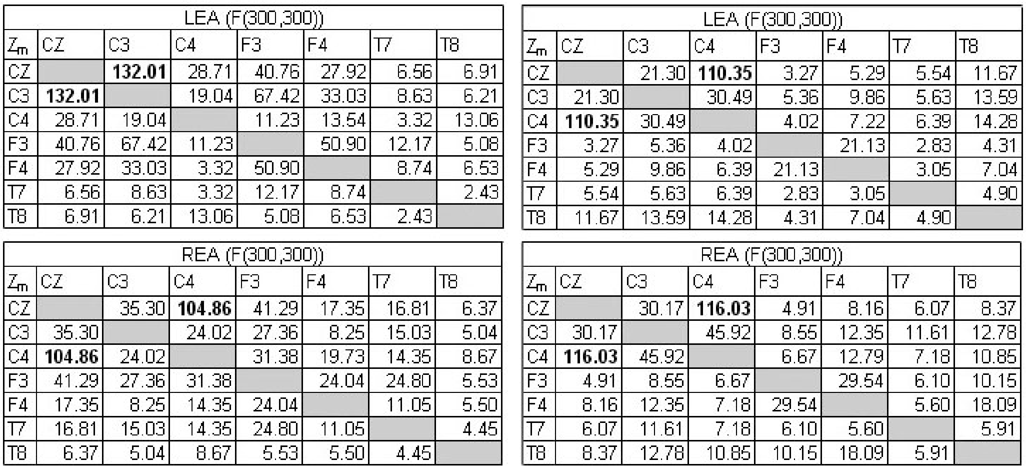


Figure 4.11 Electrodes selected to compare for similarity

time slice just before the stimulus was given. LEA, REA and HOM signals were examined individually at this stage.

Using Z_M statistic for the signals resulted as in Figure 4.12. The values greater than 1 show there is a similarity. The test statistic was compared with the value of corresponding F table value (F(300,300)) and if it is greater than the F table value, the hypothesis of *"no significance signal, just noise"* is rejected. It should be noted that greater values indicate higher similarity but not proportionally. This means that a Z_M value of 20 does not imply two times more similarity than a value of 10. The highest values of Z_M are considered in this study.

Besides electrode comparison for the signals, different sections of EEG recordings on the same electrode were also studied on similarity of REA, LEA and HOM signal averages (Figures 4.13 and4.14).



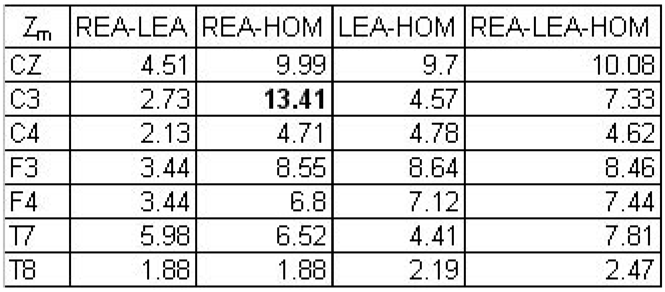
LEA (F(300,300))							
Z_m	CZ	C3	C4	F3	F4	T7	T8
CZ		**132.01**	28.71	40.76	27.92	6.56	6.91
C3	**132.01**		19.04	67.42	33.03	8.63	6.21
C4	28.71	19.04		11.23	13.54	3.32	13.06
F3	40.76	67.42	11.23		50.90	12.17	5.08
F4	27.92	33.03	3.32	50.90		8.74	6.53
T7	6.56	8.63	3.32	12.17	8.74		2.43
T8	6.91	6.21	13.06	5.08	6.53	2.43	

LEA (F(300,300))							
Z_m	CZ	C3	C4	F3	F4	T7	T8
CZ		21.30	**110.35**	3.27	5.29	5.54	11.67
C3	21.30		30.49	5.36	9.86	5.63	13.59
C4	**110.35**	30.49		4.02	7.22	6.39	14.28
F3	3.27	5.36	4.02		21.13	2.83	4.31
F4	5.29	9.86	6.39	21.13		3.05	7.04
T7	5.54	5.63	6.39	2.83	3.05		4.90
T8	11.67	13.59	14.28	4.31	7.04	4.90	

REA (F(300,300))							
Z_m	CZ	C3	C4	F3	F4	T7	T8
CZ		35.30	**104.86**	41.29	17.35	16.81	6.37
C3	35.30		24.02	27.36	8.25	15.03	5.04
C4	**104.86**	24.02		31.38	19.73	14.35	8.67
F3	41.29	27.36	31.38		24.04	24.80	5.53
F4	17.35	8.25	14.35	24.04		11.05	5.50
T7	16.81	15.03	14.35	24.80	11.05		4.45
T8	6.37	5.04	8.67	5.53	5.50	4.45	

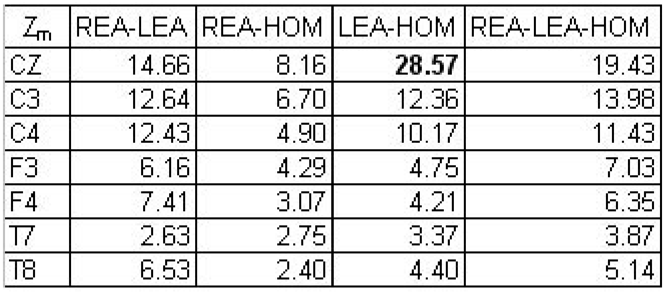
REA (F(300,300))							
Z_m	CZ	C3	C4	F3	F4	T7	T8
CZ		30.17	**116.03**	4.91	8.16	6.07	8.37
C3	30.17		45.92	8.55	12.35	11.61	12.78
C4	**116.03**	45.92		6.67	12.79	7.18	10.85
F3	4.91	8.55	6.67		29.54	6.10	10.15
F4	8.16	12.35	7.18	29.54		5.60	18.09
T7	6.07	11.61	7.18	6.10	5.60		5.91
T8	8.37	12.78	10.85	10.15	18.09	5.91	

Figure 4.12 Z_M values of electrodes in Data Set 1 and Data Set 2



Z_m	REA-LEA	REA-HOM	LEA-HOM	REA-LEA-HOM
CZ	4.51	9.99	9.7	10.08
C3	2.73	**13.41**	4.57	7.33
C4	2.13	4.71	4.78	4.62
F3	3.44	8.55	8.64	8.46
F4	3.44	6.8	7.12	7.44
T7	5.98	6.52	4.41	7.81
T8	1.88	1.88	2.19	2.47

Figure 4.13 Z_M values of REA, LEA and HOM in Data Set 1



Z_m	REA-LEA	REA-HOM	LEA-HOM	REA-LEA-HOM
CZ	14.66	8.16	**28.57**	19.43
C3	12.64	6.70	12.36	13.98
C4	12.43	4.90	10.17	11.43
F3	6.16	4.29	4.75	7.03
F4	7.41	3.07	4.21	6.35
T7	2.63	2.75	3.37	3.87
T8	6.53	2.40	4.40	5.14

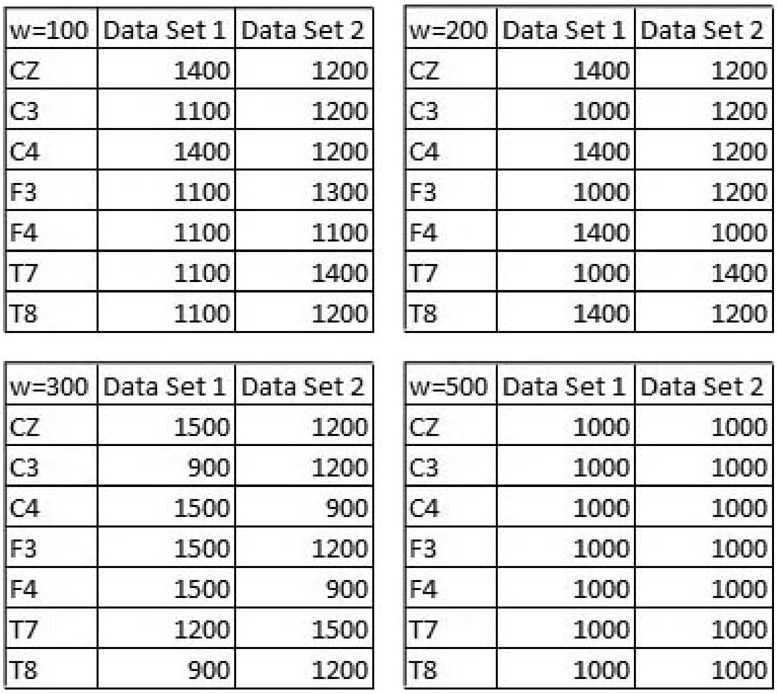
Figure 4.14 Z_M values of REA, LEA and HOM in Data Set 2

The significant observation can be seen that Data Set 1 (with Left Ear Advantage) has the greatest similarity value in REA-HOM signals on C3 electrode , but Data Set 2 (with Right Ear Advantage) has the greatest similarity value in LEA-HOM signals on CZ electrode. Apparently REA-LEA similarities on Data Set 2 are significantly greater than Data Set 1.

4.6 Most Similar Time Slices

Z_M statistic was used to find the most similar time regions of REA and LEA signals in EEG recordings. In this section of the study, [-1000;+1000] time interval was used for the analysis (0 is the time of stimuli). Each region was divided into parts of width 100 msec. Corresponding windows of REA and LEA on different electrodes were compared for similarity to detect when the most similarity is observed. The window width was updated by 200, 300 and 500 msec. and the result are given in Figure 4.15.

In large width windows most similar slices are all the same which is 1000-1500 msec. (first 500 msec. of the stimuli) for all electrodes. But narrow window widths give more detailed results. For example in w=200 msec, it is obviously seen that left side electrodes (C3, F3, T7) have a pattern of similarity in 1400-1600 msec. (400-600 msec after the stimuli) time slices for Data Set 1.



w=100	Data Set 1	Data Set 2
CZ	1400	1200
C3	1100	1200
C4	1400	1200
F3	1100	1300
F4	1100	1100
T7	1100	1400
T8	1100	1200

w=200	Data Set 1	Data Set 2
CZ	1400	1200
C3	1000	1200
C4	1400	1200
F3	1000	1200
F4	1400	1000
T7	1000	1400
T8	1400	1200

w=300	Data Set 1	Data Set 2
CZ	1500	1200
C3	900	1200
C4	1500	900
F3	1500	1200
F4	1500	900
T7	1200	1500
T8	900	1200

w=500	Data Set 1	Data Set 2
CZ	1000	1000
C3	1000	1000
C4	1000	1000
F3	1000	1000
F4	1000	1000
T7	1000	1000
T8	1000	1000

Figure 4.15 Most similar time slices of REA and LEA on different electrodes with different window sizes (w=100, 200, 300 and 500)

It was observed that Left Ear Advantage catches the similarity between left and right responses earlier in the left side electrodes (C3, F3, T7) than the right side electrodes (C4, F4, T8). But Right Ear Advantage generally has a stable similarity time in left and right responses. The greatest similarity was obtained especially in window 1000 - 1200 msec. which the stimuli is given. As the window size increases, the similarity of signal has the greatest values in the time slice that the stimuli is given (1000 - 1500 msec. for w=500).

4.6.1 Signal Similarity in Signal Shape

Z_M statistic is powerful in detecting the signal similarities in amplitude. The amplitude changes in two signals can be compared by this statistic. But, two signals may have the same shape - in different amplitudes. In this case Z_M statistic is not so powerful in detecting similarity.

In order to figure this out, a section of the EEG signal was randomly selected and a new signal was generated using this signal by adding a fixed value. To say in terms, Let $x(n)$ be a vector of any signal and let $y(n) = x(n) + 5$. In fact these are the vectors with different amplitudes but with same shape or behaviour as seen in Figure 4.16.

When Z_M statistic is calculated for $x(n)$ and $y(n)$, we compute the value of 0.4246 which is less than table value of 1. This is commented as *there is no significance signal, so these are not similar*. This is because that Z_M test statistic relies on the amplitude values of the signals. Averages and standard deviations are in concern of this statistic. So this is a disadvantage of the method.

To avoid this deficiency and to catch the behavioural similarity in signals using Z_M method, the signals were transformed into a *difference vector* computing the difference between the signal received in time *t+1* and *t*. So for a signal x, a new signal z was obtained by applying $z(t) = x(t + 1) - x(t)$. After transforming the

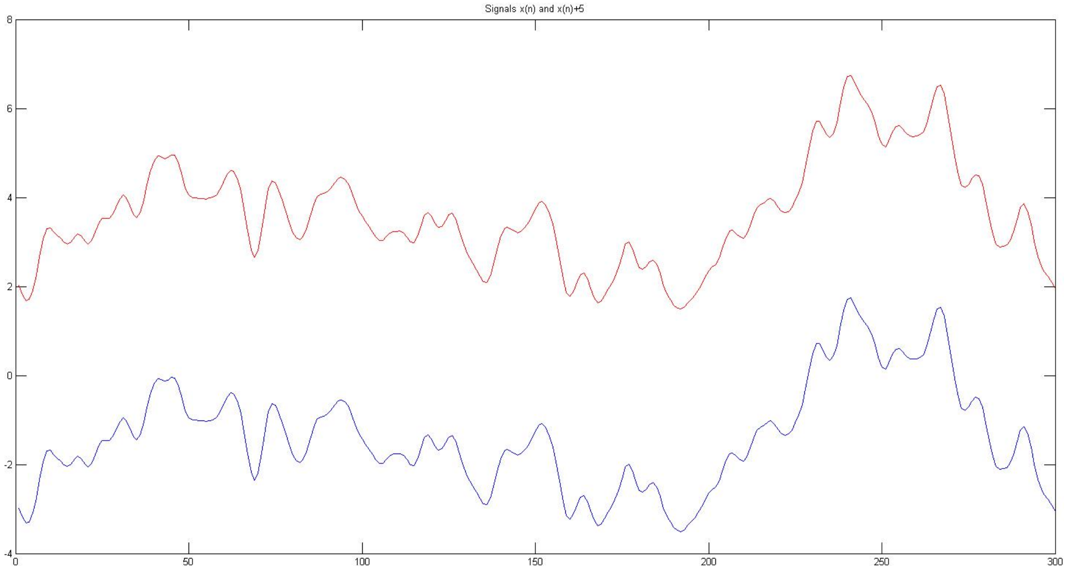


Figure 4.16 $x(n)$ and $x(n) + 5$ signals

signals to be compared, $Z_{M(diff)}$ was calculated on these difference vectors. When this transformation was applied to the previous example signal $x(n)$ and $x(n) + 5$, the Z_M value was calculated as 8.5400e+014 instead of 0.4246 which is highly less than the real similarity measure.

Using this method, presented more similarities on similar shaped signals and also displayed that some *similar* signals are *not so similar* in fact and vice versa. Examples of the mentioned situations are given in Figures 4.17 and 4.18.

This modification of Z_M was also applied to most similar time slices in order to detect a better matching of the stimulus effect. The outcomes of this modification can be seen obviously in Figure 4.19.

As a summary, Z_M statistic was applied to the difference vectors and the results were used to explain the behaviour of the two signals. Applying Z_M to the original signal helps us to comment on the similarity of the signals amplitude on the same direction, but using difference vectors explains the rate of change in amplitude and direction. This may be an advantage in EEG data. Because of the location of electrodes on the skull, some conductivity problems may occur. Some electrodes may receive the signal values weakly. In this case, using difference vector will give better results in similarity.

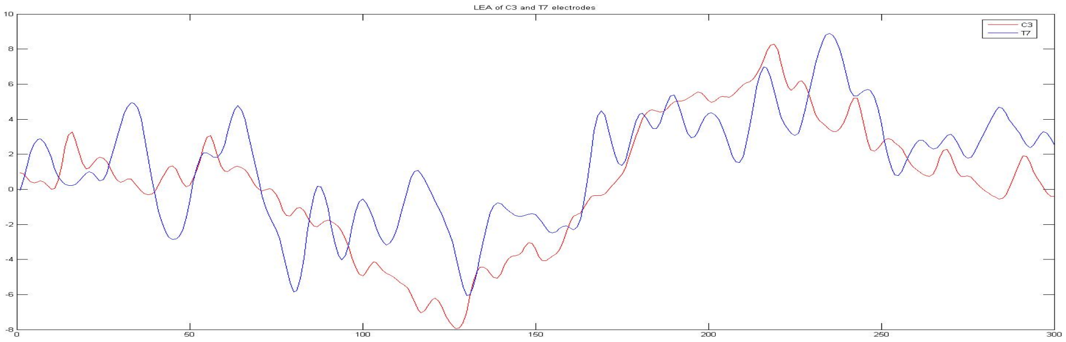


Figure 4.17 $Z_M = 5.63$ and $Z_{M(diff)} = 0.81$

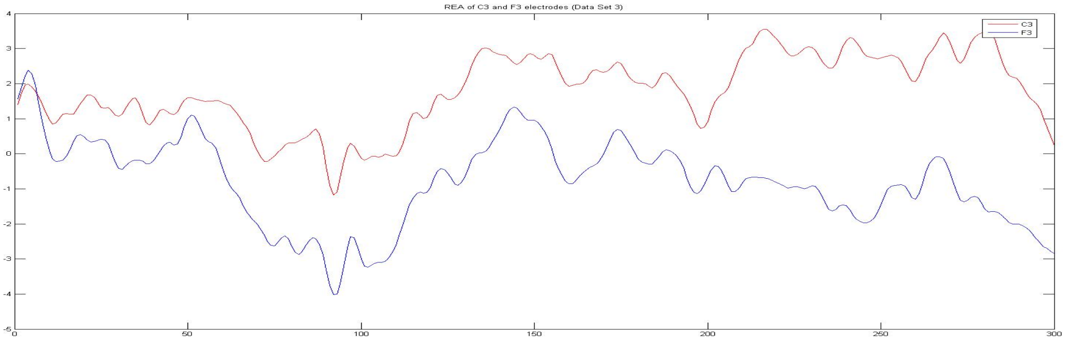
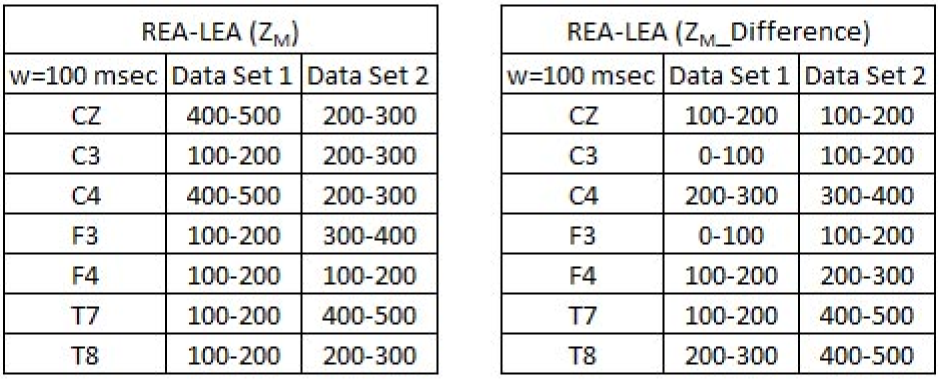


Figure 4.18 $Z_M = 0.57$ and $Z_{M(diff)} = 4.76$



REA-LEA (Z_M)		
w=100 msec	Data Set 1	Data Set 2
CZ	400-500	200-300
C3	100-200	200-300
C4	400-500	200-300
F3	100-200	300-400
F4	100-200	100-200
T7	100-200	400-500
T8	100-200	200-300

REA-LEA (Z_M_Difference)		
w=100 msec	Data Set 1	Data Set 2
CZ	100-200	100-200
C3	0-100	100-200
C4	200-300	300-400
F3	0-100	100-200
F4	100-200	200-300
T7	100-200	400-500
T8	200-300	400-500

Figure 4.19 Most similar time slices of REA and LEA on different electrodes detected with Z_M on difference vectors

4.6.2 Clustering Electrodes

In statistics, hierarchical clustering is a method of cluster analysis which seeks to build a hierarchy of clusters. Strategies for hierarchical clustering generally fall into two types:

- **Agglomerative:** This is a “bottom up” approach: each observation starts in its own cluster, and pairs of clusters are merged as one moves up the hierarchy.
- **Divisive:** This is a “top down” approach: all observations start in one cluster, and splits are performed recursively as one moves down the hierarchy.

Several criteria for determining distance between arbitrary clusters A and B is describe as follows in Larose (2005):

- Single linkage, sometimes termed the nearest-neighbor approach, is based on the minimum distance between any record in cluster A and any record in cluster B. In other words, cluster similarity is based on the similarity of the most similar members from each cluster. Single linkage tends to form long, slender clusters, which may sometimes lead to heterogeneous records being clustered together.
- Complete linkage, sometimes termed the farthest-neighbor approach, is based on the maximum distance between any record in cluster A and any record in cluster B. In other words, cluster similarity is based on the similarity of the most dissimilar members from each cluster. Complete-linkage tends to form more compact, spherelike clusters, with all records in a cluster within a given diameter of all other records.
- Average linkage is designed to reduce the dependence of the cluster-linkage criterion on extreme values, such as the most similar or dissimilar records. In average linkage, the criterion is the average distance of all the records in

cluster A from all the records in cluster B. The resulting clusters tend to have approximately equal within-cluster variability.

Using the information above, the electrodes were tried to be clustered. For the distances of the signals, correlation coefficient was used. First the correlation matrix was constructed, then the distance between two signals in two electrodes was calculated as $d = 1 - |r|$

According to this distances, average linkage clustering was applied and dendrograms for Left and Right signals were obtained as follows:

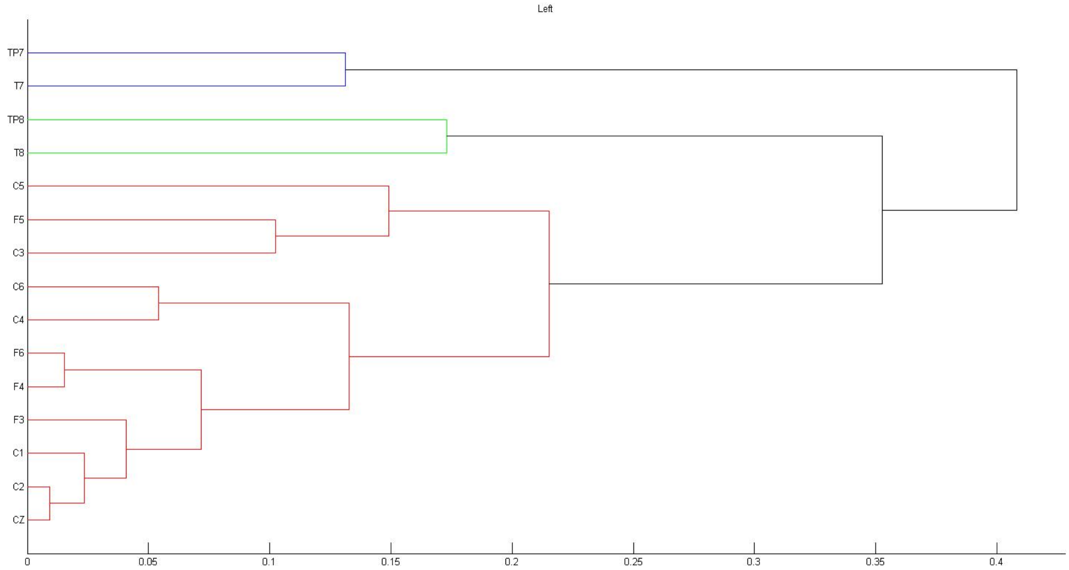


Figure 4.20 Dendrogram of Left responses

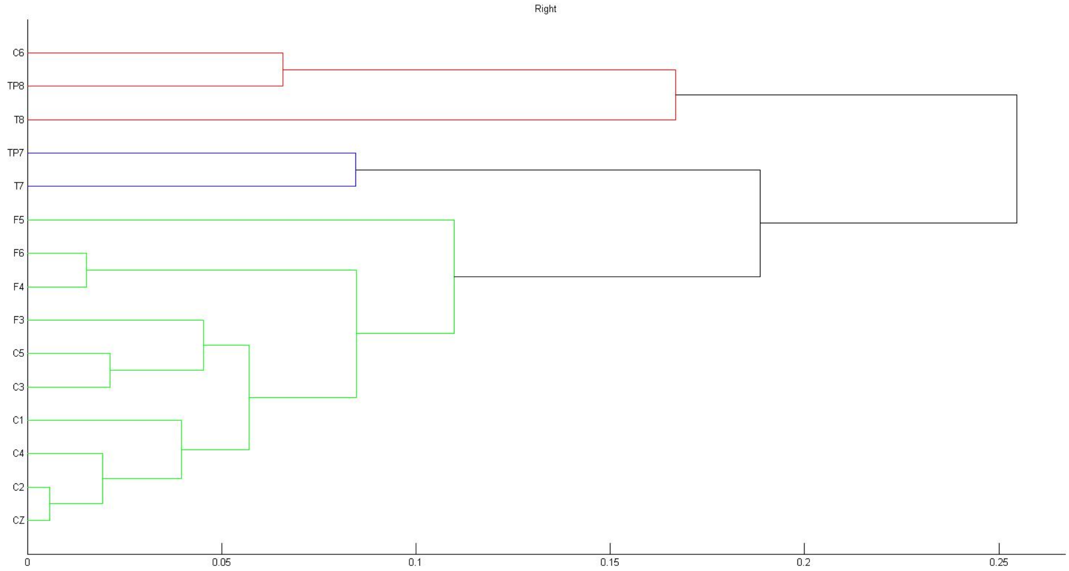


Figure 4.21 Dendrogram of Right responses

From the dendrogram, it can be commented that first joining electrodes are more similar (or correlated) than the others. Left and Right clusters were constructed differently. First joining electrodes were CZ-C2 and F4-F6. This approach may help in localising the brain in terms of electrodes in dichotic listening.

CHAPTER FIVE
DISCUSSIONS AND CONCLUSIONS

In this study of thesis, biomedical signals were examined and investigated briefly. The outcomes of body functions and especially signals received from different sections of body were studied (Rangayyan, 2002, Sörnmo & Laguna, 2005). Difficulties and interferences of biomedical signals were determined. Short introductions were given for the signals.

There are various types of imaging brain data and each has advantage or different aspects of analysing the data comparing to each other (Ray & Oathes, 2003, Britannica, 2008, Encyclopaedia Britannica, 2012, Megalooikonomou et al., 2000, Demitri, 2007). EEG was selected to study with data mining because of its advantages in size and complexity (Flexer, 2000).

In the thesis, EEG data recorded during a dichotic test were examined in detail. The data contain voltage values received in each millisecond via a cap of 64 electrodes (Bayazıt, 2009, Bayazıt et al., 2009). For a data understanding, some codes in MATLAB were written to define the related regions of the data. The EEG signals were divided into 3 partitions around the stimuli. First part is the time until the auditory dichotic stimulus is given (pre_stimuli), second part is the time between the stimulus and led indicator that tells the subject he/she can respond now (this is 2170 msec of time inteval and has fixed width for all stimuli, labelled as post_early). Third part of the signal is the interval from the led indicator until the next stimulus (post_late). Signals recorded for each stimulus and response were extracted from the whole data.Some data summarization (like answering time averages, frequencies of answers, etc.) were computed. Stimuli and responses were compared and the ear advantages of the subject were defined according to the number of answers given by each side (left or right). The responses with left ear and right ear were compared and visualised in graphics.

The signals were grouped by the answers given and average voltage values were computed for two of the datasets. It was observed that data set 1 with left ear advantage has negative averages while data set 2 with right ear advantage has positive averages. The same work was repeated for the REA, LEA and HOM signal averages and the results were presented.

In the next step, the syllables were labelled as HARD (ka, pa, ta) and SOFT (ba, da, ga) and the answers were examined under this consideration. The results differed in HARD-SOFT and SOFT-HARD syllable pairs in two data sets. Data set 2 (which has right ear advantage) is more receptive for SOFT syllable in right ear than data set 1 with left ear advantage. The distribution of answers in HARD and SOFT syllables were presented via tables and graphics. The ear advantage or brain asymmetry effect for the results was left as the topic of another study of discussion.

For the comparison of signals, the statistical similarity measure Z_M $-$ introduced in Kennedy (2007) $-$ was used to detect the similarity. Among the similarity methods mentioned in Moon (1996) and Kennedy (2007), cross correlation and Z_M statistic was applied to EEG signals to find the similarity between the signals. Z_M was selected as the primary similarity measure because Z_M statistic gave more variable results. The reason is that signals examined generally show the similar behaviour in the predefined region in terms of correlation coefficients.

Left and right advantaged answered signals were compared with each other. Different electrodes were examined for similarity. Signals between 0 and 300 msecs (where 0 is the time that stimuli is given) were taken in concern. CZ, C3, C4, F3, F4, T7 and T8 electrodes were in center of the study because the locations of these electrodes take place near to the speech and hearing functional areas of the brain (Bayazıt, 2009). The signal values were averaged grouping according to the ear advantages. The right ear answers (REA), left ear answers (LEA) and answers to homonym stimuli (HOM) were averaged for each electrode. These averages were processed for similarity.

As a result of the study, similarities on both locations (between electrodes) and ear advantage responses (REA and LEA) were detected. It was observed that similarities on REA and LEA are greater than HOM of different electrodes.

In data set 1 and data set 2, greatest similarity was detected in CZ-C4 electrodes in REA signal averages. But in LEA signals, the results differed. For the left ear advantaged data set (data set 1) CZ and C3 were most similar and for data set 2 CZ and C4 were found as most similar.

Examining the REA, LEA and HOM signals on the same electrode, it was observed that REA and HOM signals on C3 of data set 1 and LEA and HOM signals on CZ of data set 2 seem to be more similar.

As another similarity study, most similar time slices were investigated between REA and LEA signals within different electrodes. Different window widths were tried in the time interval of [-1000;+1000 msec] where 0 is the time of stimuli. It was observed that the similarities emerge after the stimulus and in the time interval of [0-500 msec]. Changing window width (100, 200, 300 and 500) resulted that Left Ear Advantage causes a similarity earlier than Right Ear Advantage on the left side electrodes (C3, F3 and T7).

The statistical measure of similarity Z_M is successful in detecting similarity in amplitude. If the signals have near average values of voltages, Z_M detects this similarity. But if the signal averages are different but their shape or behaviour is similar, then Z_M is not a reliable measure. This weakness of the method was proved by a similar example on a sample signal. The signal itself and the shifted version of the signal were compared and Z_M statistic resulted that they were not similar although they were the same signals with different voltage values.

To avoid this incapability of the method, the data were transformed into another signal by taking the difference of each data point from the previous one. The signal $x(t)$ was transformed as $z(t) = x(t+1) - x(t)$ and Z_M was applied to these new signals and the outcomes were rewarding. By doing so, Z_M method became capable to detect not only similarity in amplitude, but the similarity in behaviour or shape of the signals. The results of this and previous similarity studies showed that the modification in data using the similarity measure brought a new sight to the EEG signals. Comparing with the previous ones, some of the *dissimilar* signals appeared as highly *similar* and vice versa. This will be a great compromise because the weakness of electrode receiving small electrical signals will be confronted by this application. Besides, the effects of electrical signals within the electrodes were also eliminated in this manner. Detecting similarity in shape is more important than detecting similarity in amplitude especially in EEG signals.

At the next stage of the study, a clustering was performed over clusters. A hierarchical clustering was made and correlation coefficients of EEG signals were used to construct the distance matrix. The dendrograms given in the study presented the similarity or proximity of different electrodes in dichotic listening effects. Pre-joining electrodes were commented to be more similar than lately joining cluster.

As a sub study of thesis, entropy and wavelet topics were researched tangentially (Çek et al., 2010, Rosso et al., 2001). Entropies of auditory stimuli were analysed. Different entropy calculations and their comparisons were presented. Entropy studies on EEG data were not included in this thesis and left as further study topic.

The contributions of this study should be considered in different perspectives. This thesis is a proof of a multi disciplinary work of four branches: Computer Sciences, Statistics, Biophysics and Signal Processing. As expressed in Han & Kamber (2001), multiple disciplines confluences are obtained to get results. First year of the thesis study contains the construction of framework in order to be able to communicate within these branches. Each branch has its own vocabulary, different views for the same data and different working customs. Constructing this bridge is not a disregarded labour. Listing the contributions of this study:

- This study is a proof of multiple disciplines collaboration,
- This thesis is a good handbook for the beginners to data mining, EEG and dichotic listening,
- By this thesis, a new approach to dichotic listening on EEG data using statistical and data mining techniques was purposed,
- A new method of statistical similarity measure was used in EEG data as a new challenging work,
- The method was manipulated to detect the similarity in signal shapes, and used as a new technique in EEG recordings,
- Many open-ended research areas for further studies (like entropy, clustering, classification, similarity, etc.) were put out for researchers,
- A small library of software for manipulating EEG data in MATLAB was constructed. The functions and programs used in the thesis can be applied to many other data in different domains,
- Working on EEG data helped the author of the thesis in brain storming, coding ability, thinking in matrices and handling lots of numbers.

Future Work

Data mining individually has a large perspective of work to be done. Especially using EEG data in data mining gives fruitful results as many other domains. For the further study of this thesis, topics were listed below for exploring EEG data as a preliminary step of data mining process:

- Cleansing of EEG data is a hard and specific work area of signal processing studies. Considering that even eye lid movements effect EEG recordings, filtering or noise reduction can be a work area.
- Researching EEG data in different perspectives will be useful for obtaining different and interesting results. Not only time domain but frequency domain is a large area of work for EEG. Different frequency bands can be investigated separately. Especially in localisation studies, descriptive methods like clustering or association rules will give precious outcomes.
- Different visual presentation techniques (like graphs, charts, tables, animations, etc) will help EEG data to be better understood. Some presentations of EEG recording or dichotic listening can be prepared for researchers who are far from or afraid of this topic to be a warm up process for further studies.
- MATLAB is a good tool for EEG data handling, so some packages (like EEG ToolBox of MATLAB) can be developed or some signal processing tools can be regulated for EEG data manipulation.

Signal analysis is also another world for studying. Considering EEG signals' complexity, the researchers must be encouraged to know the basics of signal processing. This thesis focuses on signal similarity, and especially in Z_M method for measuring similarity. Besides the work done in this thesis, signal similarity offers are:

- Comparing more than two signals or electrodes may be in concern. Combination of localised electrodes particularly will give more precise results. In this thesis, binary comparisons were made. This increases the disadvantage of effects of electrodes with each other. Instead of two electrode comparisons, two groups of signals received from nearly located electrodes will be more helpful to identify the ear advantage or asymmetry in the brain.

- Searching similarities in different sections of EEG recordings may be meaningful. In this study, the EEG signals were extracted as sections like pre stimulus, post early stimulus and post late stimulus. The studies in this thesis focused on post early stimulus where the first effects of event occurred. It is known that a similar effect is observed in later periods of the signal.

- Ear and stimulus effects can be analysed particularly. A detailed analysis based on syllables or different auditory stimuli may produce interesting outcomes.

- In this thesis EEG recordings obtained during a dichotic listening test were used. Same studies can be applied for other EEG recordings such as during an epileptic attack or during sleep. Different stages of sleep can be compared in terms of similarity. Instead of auditory stimuli, visual or somatic stimuli can be used and these can be compared with each other.

- Instead of Z_M, different signal similarity methods can be used or developed. Used methods can be adjusted to be suitable for EEG signals considering the complexity and noisy structure of it.

- Searching similarity in different subject can be an interesting work. Analysing different subjects' EEG signals, particular patterns can be detected for specific stimuli.

Clustering is a giant step for data mining process. With huge amounts of data, clusters will be leading for different analysis on different targets. In this study hierarchical clustering was performed among signals received from certain electrodes. To define the *distances* of signals, correlation coefficient was used. The results were interpreted in terms of similarity again. Firstly joining electrodes were said to be more similar. The future work topics for cluster analysis can be listed as below:

- Different clustering techniques (like k-means, DBSCAN, Fuzzy c-means, etc) can be used for all of 64 electrodes to define the related localisations. It will be a good work to analyse the clusters before working on raw signals. Different clustering methods will explore the really related electrodes to work on.
- In constructing clusters, average linkage method was used in this study. Other methods like single linkage and complete linkage methods can be applied and the results can be compared.
- For the construction of distance matrix, correlation coefficient was used in this study. Various distance metrics (Euclidean, Manhattan, etc.) can be computed on signals or new distance functions can be developed.
- Signal similarity measure can also be arranged to serve as a distance measure. Applying suitable transformation to Z_M statistic or any other measure, different clusters can be obtained.
- Like clustering methods, various techniques can serve for the same purpose. Principal Component Analysis (PCA) or Independent Component Analysis (ICA) are the first featured ones. Typically Discriminant Analysis, regression techniques or neural network can also be studied as a future research.

Introducing signal similarity and signal transformation, entropy was mentioned as an important topic in the study. Entropy can be combined with similarity within the previous work. Even as a similarity measure, entropy can be computed. Also in clustering, entropy values can be arranged to become a distance or similarity identifier.

Wavelet applications provide new and fruitful working areas especially in EEG signals. Using predefined wavelets, certain EEG patterns can be determined. Another study can be performed on constructing a new wavelet for particular EEG signals. For example designing a wavelet for sleep disorder can be helpful in diagnosing particular distortions in EEG data or a brain injury.

REFERENCES

Agilent, T. (2000). *The Fundamentals of Signal Analysis*. Agilent Technologies Measurement Company.

Al-Nashash, H. A., Paul, J. S., & Thakor, N. V. (2003). Wavelet entropy method for eeg analysis: Application to global brain injury. *Proceedings of the 1st International IEEE EMBS Conference on Neural Engineering*.

Bayazıt, O., Öniz, A., Güntürkün, O., Hahn, C., & Özgören, M. (2009). Dichotic listening revisited: Trial-by-trial erp analyses reveal intra- and interhemispheric differences. *Neuropsychologia*, *47*.

Bayazıt, O., Öniz, A., Güntürkün, O., & Özgören, M. (2008). Dikotik dinleme paradigması ile beyin asimetrisinin elektrofizyolojik değerlendirmesi. *New/Yeni Symposium Journal*, *46*(3).

Bayazıt, T. O. (2009). *Uyaran Parametrelerinin EEGde Dinamik Etkileri*. Ph.D. thesis, Dokuz Eylül Üniversitesi Biyofizik Anabilim Dalı.

Bramer, M. (2007). *Principles of Data Mining*. Springer Verlag London Limited, ISBN 978-1-84628-765-7.

Britannica, E. (2008). *Britannica Guide to the Brain*. Encyclopaedia Britannica, Inc., ISBN 9781845298036.

Çek, E., Özgören, M., & Savacı, A. (2010). Continuous time wavelet entropy of auditory evoked potentials. *Computers in Biology and Medicine*, *40*(1), 90–96.

Demitri, M. (2007). *Types of Brain Imaging Techniques, Psych Central*. Retrieved February 2012 from. http://psychcentral.com/lib/2007/types-of-brain-imaging-techniques/.

Encyclopaedia Britannica, O. A. E. (2012). *Computed Tomography (CT)*. Retrieved February 2012 from.

http://www.britannica.com/EBchecked/topic/130695/computed-tomography-CT.

Flexer, A. (2000). Data mining and electroencephalography. *Statistical Methods in Medical Research*, (9), 395 – 413.

Freund, J. E. (1992). *Mathematical Statistics*. Prentice Hall, ISBN 0-13-565185-9.

Gan, G., Ma, C., & Wu, J. (2007). *Data Clustering: Theory, Algorithms, and Applications*. SIAM, Society for Industrial and Applied Mathematics, ISBN 0898716233.

Grabmaier, J., & Rudolph, A. (2002). Techniques of cluster algorithms in data mining. *Data Mining and Knowledge Discovery*, *6*, 303–360.

Han, J., & Kamber, M. (2001). *Data Mining - Concepts and Techniques*. Morgan Kaufmann Academic Press, ISBN 1-55860-489-8.

Hartigan, J. (1975). *Clustering Algorithms*. Wiley.

Hugdahl, K. (2005). Symmetry and asymmetry in the human brain. *Academia Europaea, European Review*, *13*(2).

Kennedy, H. L. (2007). A new statistical measure of signal similarity. *Information, Decision and Control - IEEE*.

Kent, R. D. (2003). *MIT Encyclopedia of Communication Disorders*. MIT Press, ISBN 9780262112789.

Kotsiantis, S. B. (2007). Supervised machine learning: A review of classification techniques. *Informatica*, *31*(3), 249–268.

Kotsiantis, S. B., Kanellopoulos, D., & Pintelas, P. E. (2006). Data preprocessing for supervised learning. *Proceedings of World Academy of Science, Engineering and Technology*, *12*.

Larose, D. (2005). *Discovery Knowledge in Data*. Wiley and Sons Inc., ISBN 0-471-66657-2.

Malmivuo, J., & Plonsey, R. (1995). *Bioelectromagnetism - Principles and Applications of Bioelectric and Biomagnetic Fields*. Oxford University Press, New York.

Megalooikonomou, V., Ford, J., Shen, L., & Makedon, F. (2000). Data mining in brain imaging. *Statistical Methods in Medical Research*, (9), 359 – 394.

Moon, T. K. (1996). Similarity methods in signal processing. *Transactions on Signal Processing - IEEE*, *44*(4).

Öniz, A. (2006). *Beyinde Delta, Teta ve Alfa Osilasyon Yanıtlarının Işığında Öğrenme Süreçleri*. Ph.D. thesis, Dokuz Eylül Üniversitesi Biyofizik Anabilim Dalı.

Polkar, R. (2001). *The Wavelet Tutorial*. Retrieved November 2009 from. http://users.rowan.edu/ polikar/WAVELETS/WTtutorial.html.

Project, C.-D. (2005). *CRoss Industry Standard Process for Data Mining*. Retrieved January 2008 from. http://www.crisp-dm.org.

Rangayyan, R. M. (2002). *Biomedical Signal Analysis*. Wiley and Sons Inc., ISBN 0-471-20811-6.

Ray, W. J., & Oathes, D. (2003). Brain imaging techniques. *The International Journal of Clinical and Experimental Hypnosis*, *51*(2), 97 – 104.

Rosso, O. A., Blanco, S., Yordanova, J., Kolev, V., Figliola, A., Schürmann, M., & Başar, E. (2001). Wavelet entropy: a new tool for analysis of short duration brain electrical signals. *Journal of Neuroscience Methods - Elsevier*, *105*.

Scientist, N. (2005). *How the Human Brain Works*. Retrieved from. http://www.newscientist.com/movie/brain-interactive.

Sörnmo, L., & Laguna, P. (2005). *Bioelectrical Signal Processing in Cardiac and Neurological Applications Biomedical Engineering*. Elsevier Academic Press, ISBN 978-0124375529.

Tan, P.-N., Steinbach, M., & Kumar, V. (2006). *Introduction to Data Mining*. Pearson Education Inc., ISBN 0-321-32136-7.

Tatum, W. O., Husain, A. M., Benbadis, S. R., & Kaplan, P. W. (2007). *Handbook of EEG Interpretation*. Demos Medical Publishing, Incorporated, ISBN 9781933864112.

Ulutagay, G. (2009). *Construction and Analysis of Clustering Algorithms Based on Fuzzy Relations and Their Applications To EEG Data*. Ph.D. thesis, Dokuz Eylül University Department of Statistics.

Vahaplar, A. (2003). *Bir Coğrafi Veri Madenciliği Uygulaması*. Master's thesis, Ege University Department of Computer Engineering.

Vahaplar, A., Çelikoğlu, C. C., & Özgören, M. (2011). Entropy in dichotic listening eeg recordings. *Mathematical and Computational Applications*, *16*(1), 43–52.

Wikipedia, T. F. E. (2006). *Electroencephalography*. Retrieved November 2011 from. http://en.wikipedia.org/wiki/Electroencephalography.

Wikipedia, T. F. E. (2009). *Electromyography*. Retrieved December 2011 from. http://en.wikipedia.org/wiki/Electromyography.

Wikipedia, T. F. E. (2012). *Brain*. Retrieved February 2012 from. http://en.wikipedia.org/wiki/Brain.

Witten, I. H., & Frank, E. (2005). *Data Mining - Practical Machine Learning Tools and Techniques*. Morgan Kaufmann Publications, ISBN 0-12-088407-0.

Zheng-you, H., Xiaoqing, C., & Guoming, L. (2006). Wavelet entropy measure definition and its application for transmission line fault detection and identificationy. *International Conference on Power System Technology*.