

1997

# ERP Analysis Using Matched Filtering, Correlation Coefficient and Mann-Whitney Test

Yun Yan  
*Portland State University*

Follow this and additional works at: [https://pdxscholar.library.pdx.edu/open\\_access\\_etds](https://pdxscholar.library.pdx.edu/open_access_etds)



Part of the [Electrical and Computer Engineering Commons](#)

Let us know how access to this document benefits you.

---

## Recommended Citation

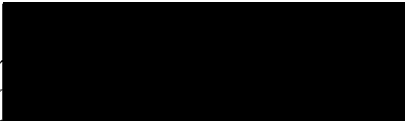
Yan, Yun, "ERP Analysis Using Matched Filtering, Correlation Coefficient and Mann-Whitney Test" (1997).  
*Dissertations and Theses*. Paper 6277.  
<https://doi.org/10.15760/etd.8137>

This Thesis is brought to you for free and open access. It has been accepted for inclusion in Dissertations and Theses by an authorized administrator of PDXScholar. Please contact us if we can make this document more accessible: [pdxscholar@pdx.edu](mailto:pdxscholar@pdx.edu).

THESIS APPROVAL

The abstract and thesis of Yun Yan for the master of Science in Electrical and Computer Engineering were presented June 5 , 1997 and accepted by the thesis committee and the department.

COMMITTEE APPROVALS:



Dr. Fu Li, Chair




Dr. Marek Perkowski



Dr. Bradford Crain


DEPARTMENT APPROVAL:



Rolf Schaumann, Chair  
Department of Electrical Engineering

\*\*\*\*\*

ACCEPTED FOR PORTLAND STATE UNIVERSITY BY THE LIBRARY

by  on 9-12-97

## ABSTRACT

An abstract of the thesis of Yun Yan for the Master of Science in Electrical and Computer Engineering presented on June 5, 1997.

Title: ERP Analysis Using Matched Filtering, Correlation Coefficient and Mann–Whitney Test

Event related potentials (ERPs) carry very important information that relate to the performance of the brain functions of a human being. Further studies have identified that the late positive complex (LPC) are affected by the memory process.

The matched filter method is used to improve the signal–to–noise ratio of signal ERPs. We use the output of the matched filter to distinguish the difference of the waveforms. In our study, we found that the peak values of the matched filter output differed among normal subjects and memory–impaired subjects.

The correlation coefficient is a statistical value that can be applied to find the degree of association between two EEG files. When there is a strong association between them, knowing one EEG file helps in predicting the other one.

A nonparametric statistical test, The Mann–Whitney Test, is introduced to set up The Filter Bank and The Correlation Bank. These two banks are very useful since the recognition percentage or correlation coefficient by the elements of the banks can distinguish whether the test subject belongs to the normal memory group or to the memory impaired group.

**ERP ANALYSIS USING MATCHED FILTERING,  
CORRELATION COEFFICIENT AND  
MANN-WHITNEY TEST**

by

YUN YAN

A thesis submitted in partial fulfillment of the  
requirements for the degree of

**MASTER OF SCIENCE  
in  
ELECTRICAL AND COMPUTER ENGINEERING**

Portland State University  
1997



## ACKNOWLEDGMENTS

The thesis presented here is not of just my effort and work. I wish to express deepest thanks to my advisor, Dr. Fu Li, for guiding me through this research with his invaluable advice, encouragement and unrelenting patience. Meanwhile, I also want to acknowledge my gratitude to Dr. Kenneth R. Erickson for his insightful guidance and sense of humor.

My sincere appreciation goes to the other members of the committee, Dr. Marek Perkowski and Dr. Bradford Crain for their helpful comments and understanding. I want to thank Hong Qi, Dianming Sun, Xueming Lin for the previous work. Thanks are also due to other members of the faculty and staff at Portland State University who have contributed greatly to my graduate education.

Finally, my parents, my husband Jim and my friend Scott, deserve my special thanks for their continued support and encouragement. I will forever be indebted to them for all that they have done.

# TABLE OF CONTENTS

<b>LIST OF TABLES</b> .....	<b>vii</b>
-----------------------------	------------

<b>LIST OF FIGURES</b> .....	<b>viii</b>
------------------------------	-------------

## CHAPTER

<b>1 INTRODUCTION</b> .....	<b>1</b>
1.1 THESIS OUTLINE .....	2
1.2 NOTATION .....	3
<b>2 EVENT RELATED POTENTIALS</b> .....	<b>7</b>
2.1 INTRODUCTION .....	7
2.2 EVENT RELATED POTENTIAL WAVEFORMS .....	8
2.2.1 ERP Waveforms .....	8
2.2.2 Electroencephalogram–EEG Signal .....	8
2.2.3 Memory Event Related Potential– –P300 .....	15
2.2.4 Visual Pattern Evoked Potentials and MS prime Task .....	19
2.2.5 EEG Data Structure .....	21
<b>3 EEG ANALYSIS USING MATCHED FILTERS</b> .....	<b>24</b>
3.1 INTRODUCTION .....	24
3.2 THE MAXIMUM SNR OF THE OUTPUT .....	25
3.2.1 The Maximum SNR of the Output .....	25
3.2.2 Test Algorithm .....	29

3.3	MATCHED FILTERING OF ERPs .....	31
3.3.1	Design of Matched Filter .....	31
3.3.2	EEG Signal Processing Using Matched Filter .....	37
3.3.3	ERP Analysis .....	40
3.3.4	Analysis Result .....	41
3.3.5	Recognition Percentages By Different Matched Filter .....	42
<b>4</b>	<b>CORRELATION COEFFICIENT &amp; MANN–WHITNEY TEST</b> .....	<b>47</b>
4.1	INTRODUCTION .....	47
4.2	CORRELATION COEFFICIENT .....	48
4.2.1	The Mean of Sample .....	48
4.2.2	The Variance and Standard Deviation of Sample .....	49
4.2.3	Sample Covariance and Correlation Coefficient .....	50
4.3	CORRELATION COEFFICIENT OF AVG FILES .....	52
4.4	MANN–WHITNEY TEST .....	60
4.4.1	Testing of Hypothesis–Statistical Inference .....	60
4.4.2	Nonparametric Test .....	63
4.4.3	Mann–Whitney Test Theory .....	64
4.4.4	Mann–Whitney Test Sample .....	67
4.5	TEST FOR CORRELATION COEFFICIENT .....	71
4.6	TEST FOR RECOGNITION PERCENTAGES .....	73
<b>5</b>	<b>COMPARISONS AND APPLICATIONS</b> .....	<b>76</b>
5.1	INTRODUCTION .....	76



5.2	COMPARISONS OF MANN–WHITNEY TEST RESULTS ...	76
5.2.1	Test Results Analysis .....	76
5.2.2	Comparison between Two Results .....	78
5.3	APPLICATIONS OF TEST RESULTS .....	78
5.3.1	Filter Bank .....	78
5.3.2	Filter Bank Means .....	79
5.3.3	Correlation Bank .....	81
5.3.4	Correlation Bank Means .....	81
5.4	COMPARISONS BETWEEN TESTS .....	83
5.4.1	Comparisons of Test Assumptions .....	83
5.4.2	Student T–test .....	84
5.4.3	Comparisons of Two Test Results .....	85
5.5	CONCLUSION .....	88
	<b>REFERENCES .....</b>	<b>90</b>

## LIST OF TABLES

TABLE	PAGE
3.1 The ERP Analysis Result Using Matched Filter MAT12 .....	42
3.2 Recognition Percentage Table for N1 N2 N3 .....	43
3.3 Recognition Percentage Table for A1, A2 and A3 .....	44
3.4 Recognition Percentage Table for JM1, JM2 and EX2 .....	45
3.5 Recognition Percentage Table for ES1, ES2, OW1, OW2 and AM1	46
4.1 Correlation Coefficient Table for MA12, N1, N2 and N3 .....	53
4.2 Correlation Coefficient Table for A1, A2, A3, JM1 and JM2 .....	54
4.3 Correlation Coefficient Table for OW1, OW2, JS1 and JS2 .....	55
4.4 Correlation Coefficient Table for AM1, AM2, EX1 and EX2 .....	56
4.5 Correlation Coefficient Table for ML1, HR1, HR2, RG1 and RG2	57
4.6 Correlation Coefficient Table for MB2, LM2, SB2, OW1 and OW2	58
4.7 Correlation Coefficient Table for MC1, MC, LS1 and LS2 .....	59
4.8 Correlation Coefficient Table for CM1, CM2, ES1 and ES2 .....	60
4.9 Annual sales volumes of college–graduated salesmen, G, and salesmen without a college degree, F. ....	68
4.10 Annual sales volumes and rank of college–graduated salesmen, G, and salesmen without a college degree, F .....	69
4.11 Mann–Whitney Test Results for Correlation Coefficient .....	72
4.12 Mann Whitney Test Results for Recognition Percentage .....	74
5.1 Filter Bank .....	76
5.2 Correlation Bank .....	77

5.3 Mean Table for The Filter Bank .....	79
5.4 Test Sheet for Recognition Percentage .....	80
5.5 Mean Table of The Correlation Bank .....	82
5.6 Test Sheet for The Correlation Bank .....	83
5.7 Comparison Table for The Filter Bank .....	86
5.8 Comparison Table for The Correlation Bank .....	87

## LIST OF FIGURES

FIGURE	PAGE
2.1 A single plane projection of the head, showing all standard positions and the location of the rolandic and sylvian fissures. ....	10
2.2 Frontal view of the skull showing the method of measurement for the central line of electrodes .....	11
2.3 EEG Raw Data (1–10000 Sampling Points) .....	13
2.4 Averaged ERP's of 20 signal EEG trials of low–tone–high–tone task. ....	14
2.5 Averaged normal ERP's of 40 single EEG trails of MSprime task.	17
2.6 Averaged memory–impaired subject ERP's of 40 single EEG trails of MSprime task .....	18
2.7 Two Segments of the MSprime task .....	20
3.1 The Matched Filter for a Physically Realizable System, minimum delay .....	28
3.2 The design procedure of the matched filter (a) .....	32
3.2 The design procedure of the matched filter (b) .....	33
3.2 The design procedure of the matched filter (c) .....	34
3.2 The design procedure of the matched filter (d) .....	35
3.2 The design procedure of the matched filter (e) .....	36
3.3 The matched filtering result of 20 sweeps of EEG signal of a normal normal memory subject .....	38

3.4 The matched filtering result of 20 sweeps of EEG signal of a memory-impaired subject .....	39
4.1 Region of rejection for testing $H_0: m = m_0$ against $H_1: m \neq m_0$ ...	62

# CHAPTER 1

## INTRODUCTION

Event related potentials (ERPs) carry very important information that relate to the performance of the brain functions of a human being. A number of studies have identified that one component, in particular, the late positive complex (LPC), is affected by the memory process. The small amplitudes of ERPs are embedded in the ongoing electroencephalogram (EEG) signal which has an amplitude several times larger than ERPs plus other background noise. Thus, the signal-to-noise ratio (SNR) is less than 1:1 (0 dB). This small SNR is one of the most difficult issues in the field of ERP analysis.

The matched filter method is used to improve the SNR of signal ERPs. We use the output of the matched filter to distinguish the difference between waveforms from normal subjects and memory-impaired subjects. The concept of **recognition percentage** for each subject is introduced as one of the two fundamental parameters by our research, which we apply for our statistical Mann-Whitney test.

Another parameter we compare while using the Mann-Whitney test is the **correlation coefficient**. Concerning the significant differences between the EEG files of all subjects, we want to calculate a value to tell how two EEG files relate to each other. The **correlation coefficient** serves as a measure of the extent to which two EEG files are dependent.

The Mann–Whitney test is a nonparametric test. From the **recognition percentages** and **correlation coefficients** of Group A (Normal Memory Subjects) and Group B (Impaired Memory Subjects), we apply them in the Mann–Whitney test to set up the Filter Bank and the Correlation Bank. The Filter Bank is composed of EEG files, which significant level of **recognition percentage** is under 0.05 by Mann–Whitney test. The Correlation Bank consists of EEG files, which significant level of **correlation coefficient** is under 0.05 by Mann–Whitney test.

The last step of this research is to apply both the Filter Bank and the Correlation Bank to the clinical analysis. First, we use each of the averaged EEG files by the sample points in the Filter Bank as a filter to calculate the **recognition percentage** of the test subject. With the calculation result and the **recognition percentage Mean Value Table** for the Filter Bank, we can evaluate the memory status of each test subject. Second, we use each EEG average file of the Correlation Bank and the EEG average file of test subject to calculate their **correlation coefficients** between them. With the calculation result and the Correlation Coefficient Mean Value Table for the Correlation Bank, we can analyze the special properties of the test subject.

## 1.1 THESIS OUTLINE

This thesis is organized as follows:

Chapter 1 : INTRODUCTION – – General introduction about this thesis.

Chapter 2 : EVENT RELATED POTENTIALS – – An introduction of EEG signals and ERPs. We focus on the component of the late positive complex (LPC) which is related to the memory process.

Chapter 3 : EEG ANALYSIS USING MATCHED FILTERS-- A review of the theory of the matched filter and the use of the ensemble averaging of EEG signals (referred as AVG files) to design a matched filter to improve the signal-to-noise ratio (SNR) of signal ERPs. After matched filtering, we analyze the ERPs by comparing the peak values of the different stimuli to get the value of the **recognition percentage**. The EEG signals of a total of 16 trials, 9 for normal subjects and 7 for memory-impaired subjects, are processed. The results show a comparatively clear pattern. The value tables of **recognition percentage** using each EEG file as a matched filter are given in this chapter.

Chapter 4 : CORRELATION COEFFICIENT and MANN-WHITNEY TEST-- A review of the theories and applications of both **correlation coefficients** and Mann-Whitney test. Several related statistics concepts are introduced. The value tables of **correlation coefficients** between the AVG files are given in this chapter. We explain why we use the Mann-Whitney test here instead of the Student T-test, and how we use it with both Recognition Percentage Tables and Correlation Coefficients Tables to get The Filter Bank and The Correlation Bank.

Chapter 5 : COMPARISONS and APPLICATIONS-- A comparison between P values of Mann-Whitney test for both **correlation coefficient** and **recognition percentage**. In the application section, we focus on how to use The Filter Bank and The Correlation Bank to analyze the test subjects. Another brief comparison between Mann-Whitney test and Student T-test.

## 1.2 NOTATION

Z denotes the set of integers.



$L^2$  denotes the Hilbert space of measurable, square-integral one dimensional functions such that

$$\int_{-\infty}^{+\infty} |f(x)|^2 dx < +\infty \quad 1-1$$

We denote the convolution of two functions  $f(x) \in L^2$  and  $g(x) \in L^2$  as

$$f(x) * g(x) = \int_{-\infty}^{+\infty} |f(u)| du * g(x - u) du \quad 1-2$$

The Fourier transform of any signal  $f(x)$  is written by  $\hat{f}(\omega)$  and is denoted by

$$\hat{f}(\omega) = \int_{-\infty}^{+\infty} f(x) * e^{-i\omega x} dx \quad 1-3$$

For any function  $f(x)$ ,  $f_s(x)$  denotes the dilation of  $f(x)$  by the scale factor  $s$

$$f_s(x) = \frac{1}{s} f\left(\frac{x}{s}\right) \quad 1-4$$

RP denotes the **recognition percentage**.

CC denotes **correlation coefficient**.

ERPs: Event related potentials (ERPs) carry very important information that relate to the performance of the brain functions of the human being. In particular, the P300, or the late positive complex (LPC), are affected by the memory process.

Matched Filter: It is a typical signal processing method to obtain the known signal from background. It yields a maximum signal-to-noise ratio when the

signal with additive noise, if the noise is the white noise, passes through it. We use output peak values from matched filter to calculate RP.

**Correlation Coefficient:** It is a statistical parameter applied to find the degree of association between two averaged electroencephalogram (EEG) files.

**Mann–Whitney Test:** It is a nonparametric statistic test. It is used here to distinguish whether there is a significant difference between two group means by both recognition percentage and correlation coefficient.

**Recognition Percentage:** It is calculated by the peak values of the matched filter output.

**EEG:** Electroencephalogram is the recording of brain electric potentials varying in time at frequencies. ERPs are embedded in EEG signal.

**AVG file:** Another most commonly used method to improve the SNR is ensemble averaging of the signal. AVG file is the averaged EEG file, which can show fairly clear pattern for normal and impaired ERP waveforms.

**Nonparametric Test:** It is a statistical procedures that doesn't require knowledge of the form of the probability distribution from which the measurements come.

**LPC:** Late Positive Complex of the long–latency components in the ERP waveforms which are affected by the memory process.

**Significant Level:** The level of significance refers to the state of being “statistically significant”. Once the level of significance is chosen the region of rejection  $\alpha$ , also called the critical region, is decided upon.

**P value:** P values report the smallest level at which the observations are significant, the level of just significance or the critical value. If the P value is

smaller than the nominal level, the observations are significant, and otherwise not significant.

For EEG signals and the corresponding outputs of the matched filter, 256 sampling points were used to sample each 1.024-second-long sweep, so each sampling point corresponds to 0.004 second.

All EEG signals, including raw data and averaged ERP waveforms used in this thesis, were provided by the Erickson Memory Clinic and Research Center.

## **CHAPTER 2**

### **EVENT RELATED POTENTIALS**

#### **2.1 INTRODUCTION**

Brain signals research, which arises from the utilization of brain signal as clinical and research tools and its contributions to the basic understanding of the functions of the brain, has been a very important research issue. People seek to elucidate the fundamental steps for the various functions of the human brain and predict the functionally relevant diseases. Brain waves provide a classic example of a non-stationary, multi-dimensional signal processing problem. Being a main research resource, ERPs (Event Related Potentials) do carry very important information, but the low SNR and the variability of the latencies and amplitudes of the components make obtaining the information of brain function very difficult [1].

In this chapter, we introduce, in general, the ERP waveforms and electroencephalogram (EEG) signals and the memory event related potential – the P300, or as it is sometimes termed, the late positive complex (LPC). A visual paradigm designed to elicit the LPC and the method to measure it and how to apply it will be explained in further chapters.

## **2.2 EVENT RELATED POTENTIAL WAVEFORMS**

### **2.2.1 ERP waveforms**

The fact has been confirmed that Event Related Potentials reflect a number of cognitive variables in a systematic manner. ERPs are elicited by the application of sensory stimuli, e.g., visual or auditory, and are of a complicated transient nature characterized by a distinct onset and finite duration.

ERPs may be especially useful for determining how much, or to what depth, processing is carried out upon relevant stimuli. The conventional approach is to model the ERP as a deterministic function for repetitive stimulations, in which case the ensemble average of a number of responses will give the best estimate when the noise is random and is of zero mean. However, there is much empirical evidence that ERP waveforms vary randomly from stimulus and therefore much interest is currently focused on single ERP waveforms.

### **2.2.2 Electroencephalogram—EEG Signal**

The Electroencephalogram (EEG) is the recording of brain electric potentials varying in time at frequencies extending up to a few tens of cycles per second and measuring from a few microvolts up to a few millivolts. They are related to important aspects of information processing in the brain. These low voltages are measured by scalp electrodes placed at various positions and amplified by an EEG amplifier, the output of which drives various recording

instruments. The raw AVG file is the averaged raw EEG file by each channel. The method of averaging EEG signals can show fairly clear pattern for normal and impaired ERP waveforms.

A number of studies have observed that the ERP waveform does not maintain a uniform shape. It differs with respect to electrodes distribution. In our study, after careful comparison, we use the data recorded from Cz [2] (nomenclature is from the International 10–20 System). See Figure 2.1 and the reference point referenced to a common ground, to be our experimented resource [3] [4]. Figure 2.2 shows a frontal view of the skull showing the method of measurement for the central line of electrodes [5] [6]. This method is designed to cover various brain regions and lobes, thus the labeling of the electrodes is in accordance with their location over brain structures.

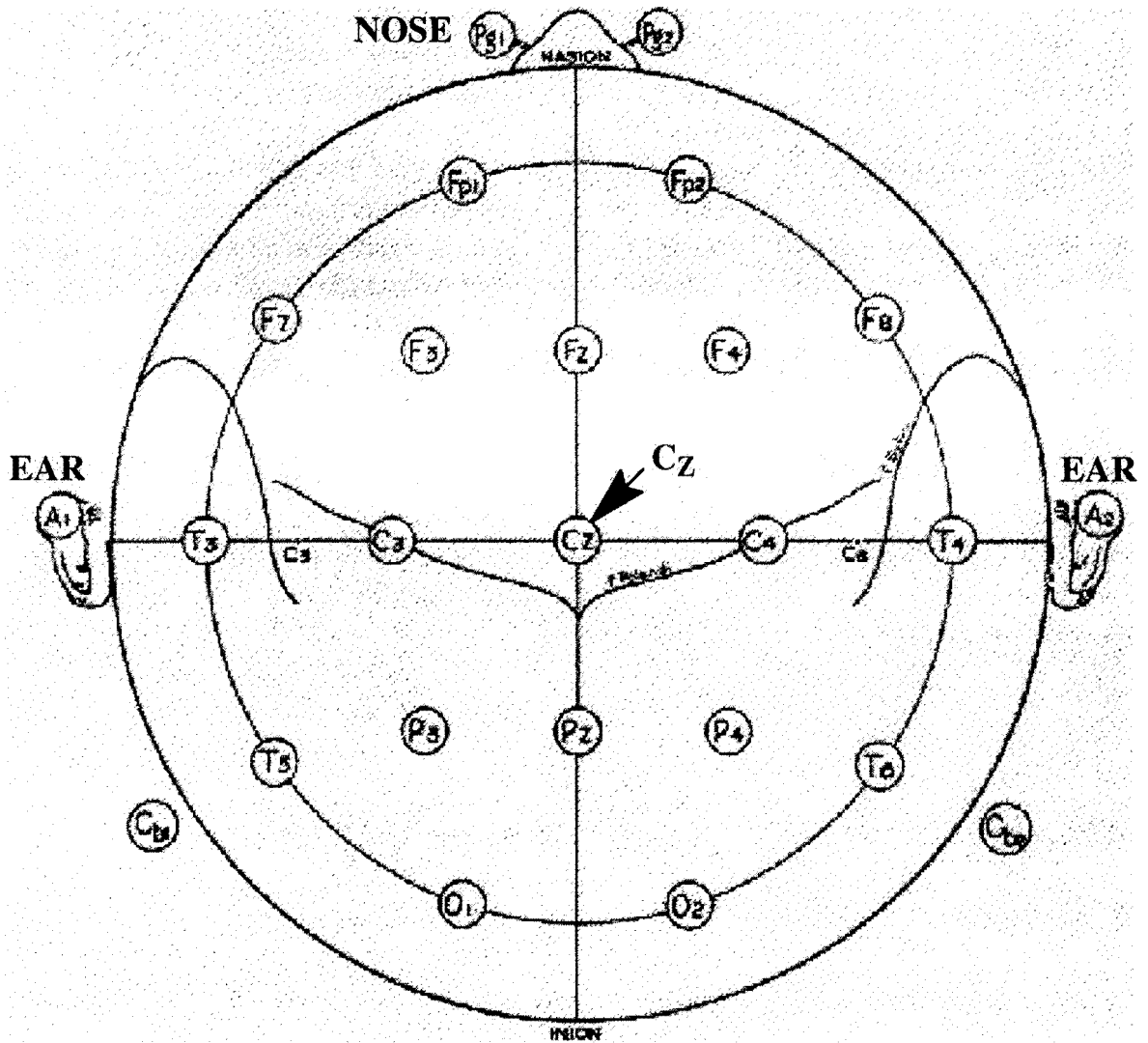


Figure 2.1 A single plane projection of the head, showing all standard positions and the location of the rolandic and sylvian fissures.

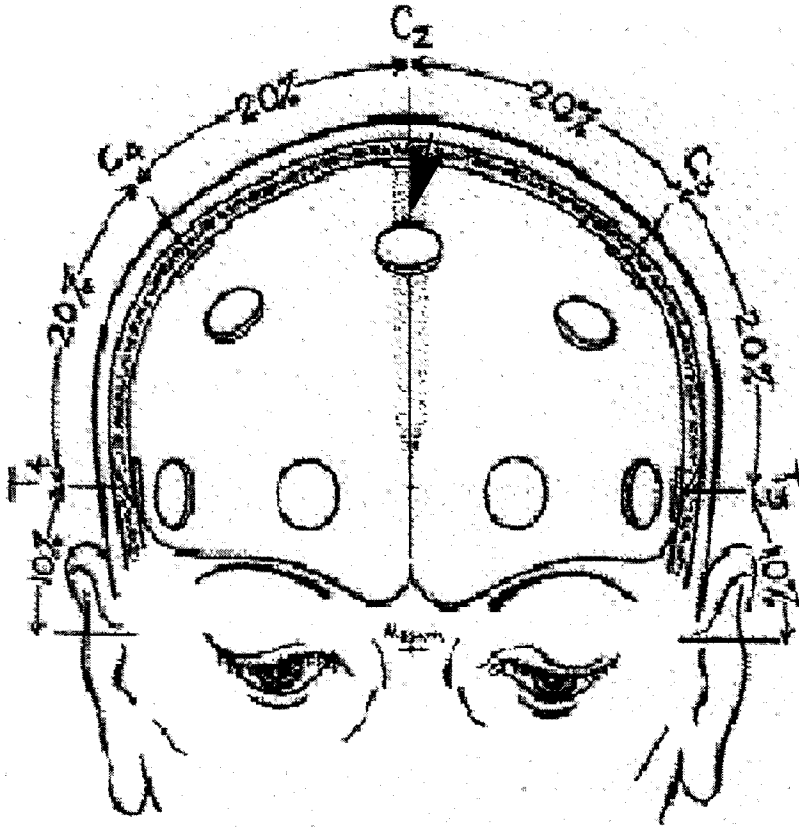


Figure 2.2 Frontal view of the skull showing the method of measurement for the central line of electrodes

ERPs are measured using one or more scalp electrodes and most generally referencing the measurements to a body position such as linked ears that is located some distance away from the area of the cortex where the response is expected. The amplitudes of the ERPs vary from tenths of a microvolt to tens of microvolts and are embedded in the ongoing EEG waveform whose amplitude is typically  $10-30 \mu\text{V}$ , which is the recording of brain electric potentials varying in time at frequencies. Thus, in many instances the signal-to-noise ratio (SNR) is less than 1:1 (0dB) [7]. It is this small SNR that makes waveform anal-



ysis difficult. A segment of raw EEG signal is shown in Figure 2.3, we can not see the ERP waveform due to the noise.

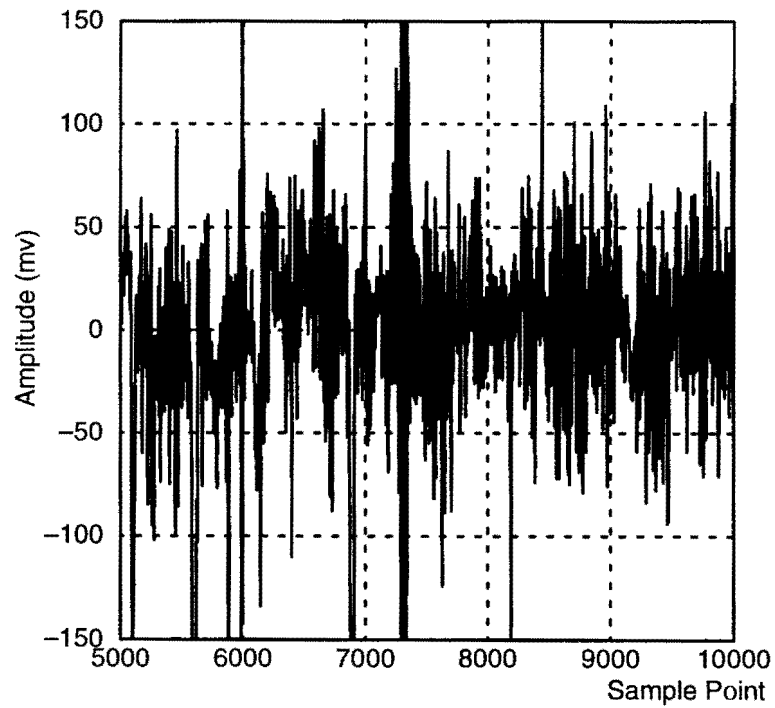
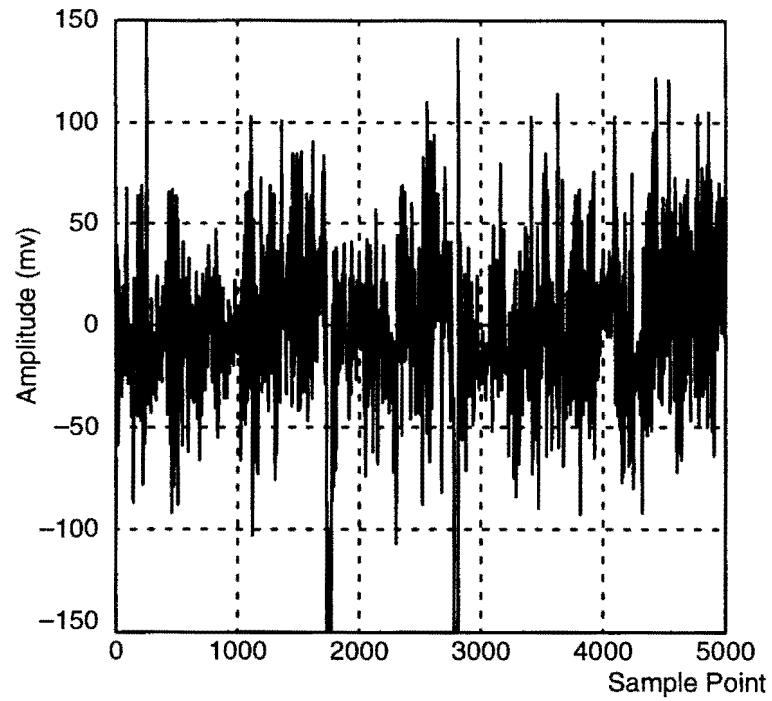


Figure 2.3 EEG Raw Data (1–10000 Sample Points)

Electrophysiologic investigations into cognitive process during the past two decades have identified certain evoked potential components (termed event-related potentials, or ERPs) which appear sensitive to psychological factors [8]. Figure 2.4 shows the ERP waveform obtained by averaging 20 single trials of the auditory task. Each latency is related to one certain event.

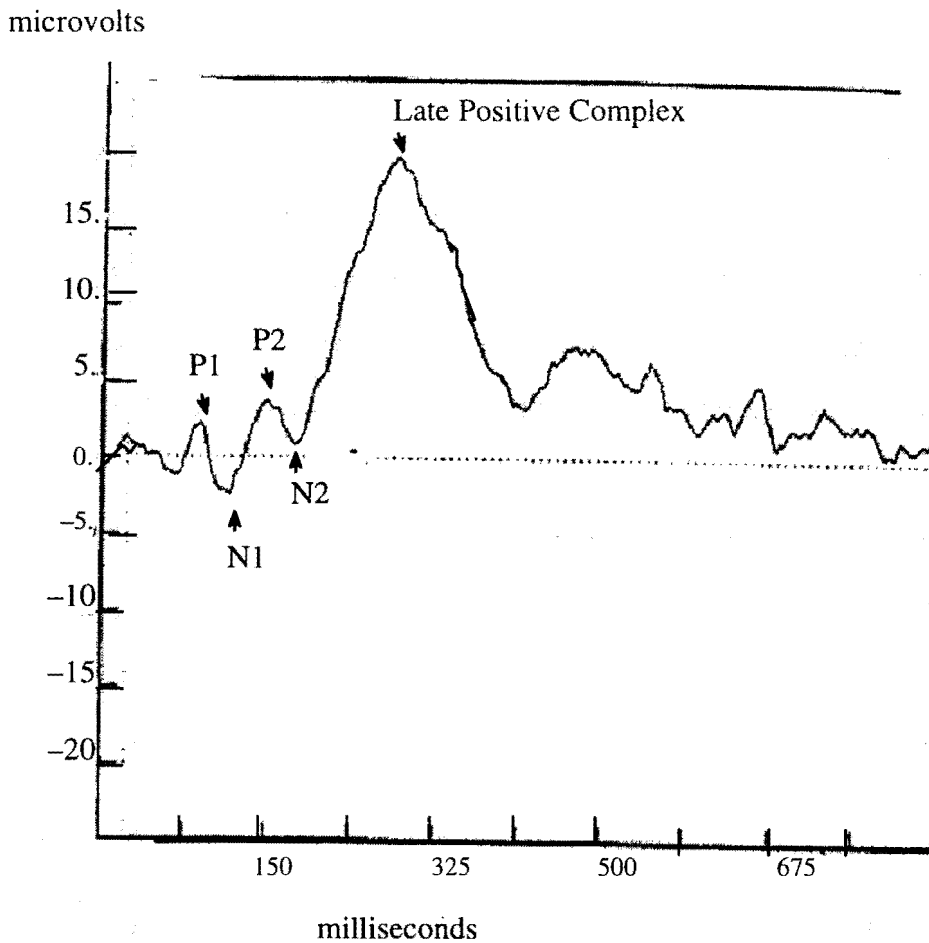


Figure 2.4 Averaged ERPs of 20 signal EEG trials of low-tone-high-tone task

### 2.2.3 Memory Event Related Potential – –LPC

Generally, a typical ERP waveform can be divided into three main segments: (1) P1–N1–P2–N2 complex. (2) Late positive complex component and (3) a late potential. An average of the 20 target trials is shown in Figure 2.3. Of chief interest among these has been the late positive complex, a positive waveform peak having maximum excursion at around 300 milliseconds following stimulus presentation.

Peak analysis is the simplest way of reducing the mass of data in an averaged waveform so that the waveform can be specified in terms of a few numbers. Because a peak value usually appears at about 300 ms after stimulus, it is also called P300. It is a very important feature in brain waveform analysis. The latency varies with the subjects attention, alertness, age, stimulus processing speed and memory ability.

The existence of an independent P300 related to specific cognition and stimulus identification is no longer the issue. P300 components have been elicited under conditions of uncertainty when the stimulus delivers feedback concerning the accuracy of a guess or of a judgment, in situations where the subject is required to make a choice response as soon after stimulus presentation as possible, and in situations where low probability targets are presented against a background of more probable nonsignals. In our research, we use MSprime task files to elicit P300, which will be discussed in detail in the next section.

The P300 represents a nonspecific reactive change of state subsequent to cognitive evaluation of significant stimuli. It is emitted when a subject recognizes an important but unexpected stimulus. Its amplitude also depends very much, however, on the likelihoods or expectancies for the different classes of stimulus and responses that may be required. Also, it is relatively independent of the particular sensory modality, and largely unaffected by stimulus parameters such as intensity, pitch, color, size, etc. Some investigators have postulated that P300 latency is a measure of the time required for such processes as stimulus evaluation and categorization. They have found its latency to be relatively independent of processes underlying response selection and execution [9].

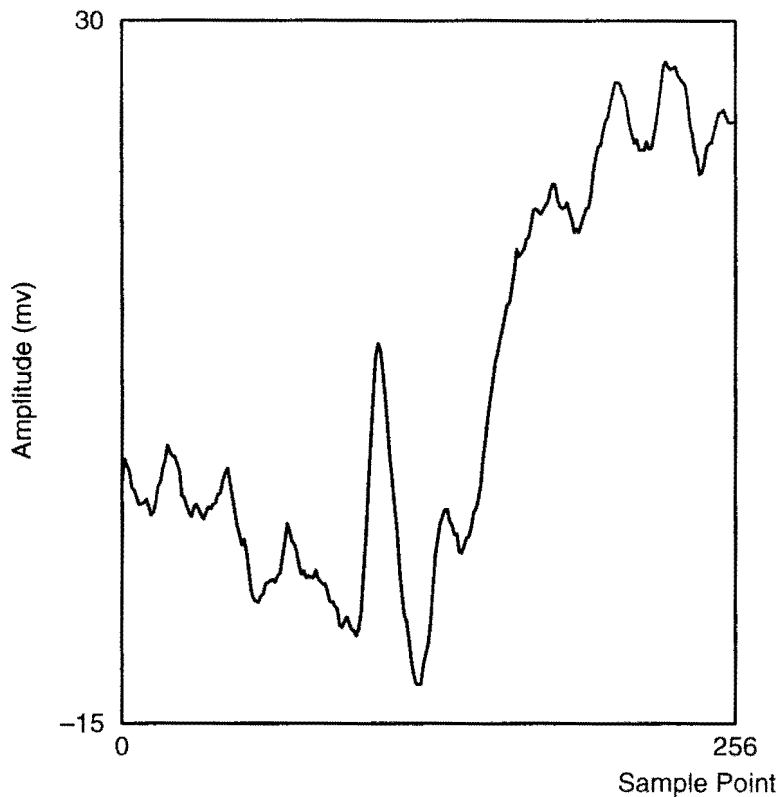


Figure 2.5 Averaged normal ERPs of 40 single EEG trails of MSprime task

The ERP components have different relationships to stimulus probability and task relevance. Numerous studies have further distinguished positive components with latencies of 400 milliseconds and beyond with various relationships to processes of stimulus categorization, and response selection and execution. Current nomenclature for the series of components with latencies from 300–600 milliseconds is the “late positive complex”. Figure 2.5 is the averaged waveform of a normal subject for MSprime task and Figure 2.6 is from a memory–impaired subject for the same task. We can easily find the differ-

ence in the late positive complex between the normal and impaired ones. But the P300 component is not clear in Figure 2.6. It is embedded in the late positive complex.

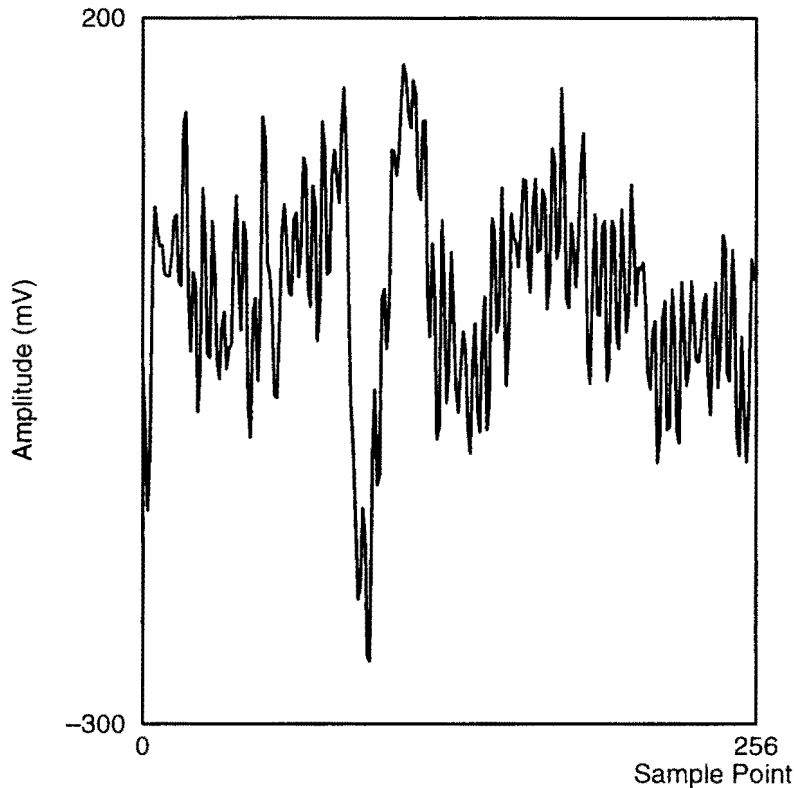


Figure 2.6 Averaged memory-impaired subject ERPs of 40 single EEG trails of MSprime task

The method of averaging EEG signals can show fairly clear patterns for normal and impaired ERP waveform. But the real challenge is to identify the waveform that varies significantly between demented and normal subjects from signal trials.

## **2.2.4 Visual Pattern Evoked Potentials and MS prime Task**

### **1. Visual Pattern Evoked Potentials**

Humans' eyes are an exquisite apparatus representing both an optical and a neuronal device. Light entering the eye must pass through transparent media: the cornea, the aqueous humor, the lens and the vitreous humor to reach the retina. The retina is a neuronal membrane lining the back of the eye chamber.

The most important human visual function is the ability to see and recognize objects. The general aim of pattern VEP studies is to advance our understanding of the sensory aspect of this visual function and to clarify its physiological basis.

Later aspects of the VEP incorporate the late positive complex (LPC) which is sensitive to processing word meaning, and match or mismatch in meaning (Fig 2.5).

### **2. MS Prime Task**

In our studies we set up a semantic context by presenting three meaningful words visually, one word at a time. After every three related words, subjects would be asked to judge if the fourth (target) word is related to the first three words. We want to examine in particular the properties of the LPC that would be elicited when the unexpected word that is semantically unrelated with the



previous three words is inserted at the end. The subjects were shown a total of 160 words on the screen. In one complete experiment, there are four groups of 40 sets. In each set we call the first three words “priming words”. The fourth word, or target word, was presented after a warning tone. The target words were randomly interspersed to be related or unrelated along the train of 40 sets. Figure 2.7 shows a example of a segment of the task.

After the first three priming words: Lady, Clinic and Coat, a beep would indicate to the subject that the target word would come next. The target word, “nurse”, is related to the three priming words. For the second set, the target word, “button”, is not related to its priming words: Indian, River, and Boat. For data analysis, we call all the priming words condition 1, the related words condition 2, and the unrelated words condition 3.

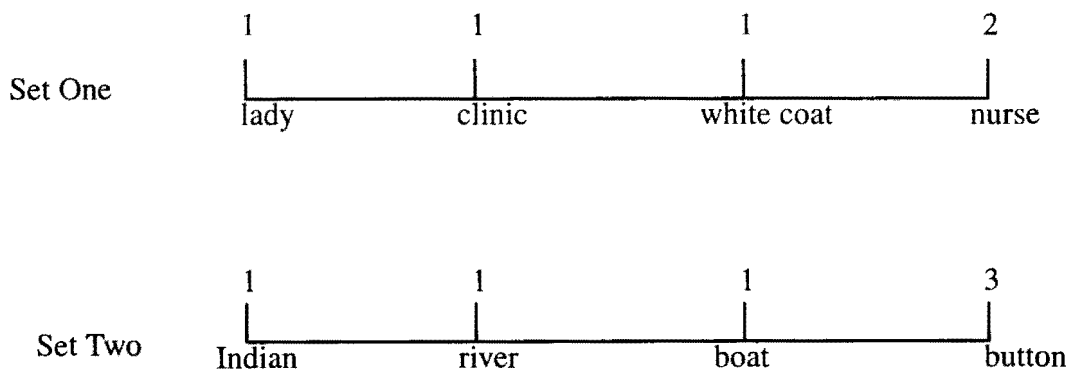


Figure 2.7 Two Segments of the MSprime task.

In one single ERP trial, a signal is recorded from 200 milliseconds before the stimulus to 824 milliseconds after the stimulus using 256 sampling points (one every 4 millisecond). We call this single segment one sweep. One hundred and sixty 1.024–second–long EEG sweeps are combined into one file.

In every set, we report only the ERP responses elicited by the first of the three priming stimuli and the corresponding target word.

### **2.2.5 EEG Data Structure**

The raw data files from the clinic have two types: one is \*.eeg, the other is \*.avg.

The structure and format of raw \*.eeg files are as follows:

1. The original data file is binary code.
2. The content of the data should be integer.
3. There are 24 channels recorded in each file, each channel has 256 time points, and there are 80 trials in one file.
4. From the very beginning of the files, comes
  - a. file header (900 bytes)
  - b. 75 bytes subheader for each channel , so together  $75 \times 24 = 1800$  bytes.
  - c. data block:

first are 13 bytes for labelling the first sweep

Coming next are data at first time point for 24 channels, each time point represented by 2 bytes (integer), low–byte appearing first, and then the high–byte (the highest bit is the sign bit).

And then are the data at the second time point for 24 channels, etc., till the 256Th time point.

Next should be 13 bytes for the second sweep, and then data for 24 channels of 256 time points, etc., till the 80 or 160 sweeps are completed.

5. So the length of the binary code should be

$$900 + 75 \times 24 + (13 + 2 \times 24 \times 256) \times 80 = 986780.$$

6. Use “eeg” to read the \*.eeg file.

The structure and format of raw \*.avg files are as follows:

1. The original data file is binary code.
2. The content of the data should be float.
3. There are 24 channels recorded in each file, each channel has 256 time points.
4. From the very beginning of the files, comes

- a. file header 900 bytes.
- b. 75 bytes subheader for each channel, so together  $75 \times 24 = 1800$  bytes.
- c. data block:

first are 5 bytes for labelling the first channel.

coming next are mean value data of 256 time–point for the first channel, each time–point data has 4 bytes(float).

following are variance value data of 256 time– point for the first channel, each time–point represented by 4 bytes (floating point).

next are the 5 bytes for the second channel, and then mean value data of 256 time–point data for the second channel, etc., till the 24th channel.

5. So, the length of the binary code should be (suppose only one sweep in one file):

$$900 + 75 \times 24 + ( 5 + 4 \times 256 + 4 \times 256 ) \times 24 = 51972$$

In our research, we only pick up Channel 4 data from the raw data files to apply for our EEG file and AVG files.

## CHAPTER 3

### EEG ANALYSIS USING MATCHED FILTERS

#### 3.1 INTRODUCTION

Extraction of Event Related Potentials (ERPs) from the background EEG is an important issue in brain research [10]. A major problem in the extraction process is the poor Signal-to-Noise Ratio (SNR), which characterizes ERPs relative to the ongoing background cerebral activity. The most commonly used method to improve the SNR is ensemble averaging of the signal, time-locked to some external trigger. Many advanced methods that are currently under investigation, apply a variety of adaptive filtering techniques aimed at reducing the number of repetitions, ideally to a single trial. Some are based on an assumed mean behavior of the underlying signal, by which they design optimal filters. Others assume a stationary model for ongoing EEG activity, and apply prewhitening techniques to the signal trials, hoping to reduce the noise with a minimal signal distortion. However, none of the suggested methods have yet become routine in brain research, due to a high complexity required for some, or only a minor signal improvement for others.

In this chapter, the Matched Filter method is introduced and we demonstrate how it is used to improve the SNR for single trial ERPs and then analyze

the ERPs. We also give the definition of the **recognition percentage** of a filter, and at the same time calculate **recognition percentages** by all filters.

## 3.2 The MAXIMUM SNR OF THE OUTPUT

### 3.2.1 The Maximum SNR of the Output

Matched Filter is a linear time-invariant filter. Matched filter yields a maximum Signal-to-Noise Ratio when the signal with additive noise [11], if the noise is white noise, passes through it. Specifically, if the noise is a Gaussian noise, then the matched-filter detector minimizes the probability of detection error when the threshold level is properly set.

When we set up a system to get its output, we are interested in maximizing the peak pulse signal in the presence of additive noise, especially, in the case in which the signal pulse additive noise is passed through a linear time-invariant filter. Out of all filters, we want to yield a maximum output.

Let the signal input to the filter be  $[f(t) + n(t)]$ , where  $f(t)$  is the signal and  $n(t)$  is the additive noise. The output of the filter is  $[f_0(t) + n_0(t)]$  and we wish to maximize the ratio  $|f_0(t_m)| / [n_0^2(t)]^{1/2}$ , where  $t = t_m$  is the best observation time ( to be set ). Actually, as we shall see , it turns out to be more convenient to maximize the square of this ratio.

Let the Fourier transform of  $f(t)$  be  $F(\omega)$  and let  $H(\omega)$  be the frequency transfer function of the desired optimum filter. Then we can write

$$f_0(t) = \frac{1}{2\pi} \int_{-\infty}^{\infty} F(\omega) H(\omega) e^{j\omega t} d\omega \quad 3-1$$

$$f_0(t_m) = \frac{1}{2\pi} \int_{-\infty}^{\infty} F(\omega) H(\omega) e^{j\omega t_m} d\omega \quad 3-2$$

The power spectral density of the noise  $S_n(\omega)$ , so that

$$\overline{n_0^2(t)} = \frac{1}{2\pi} \int_{-\infty}^{\infty} S_n(\omega) |H(\omega)|^2 e^{j\omega t_m} d\omega \quad 3-3$$

Dividing the squared magnitude of Equation 3-2 by 3-3, we get

$$\frac{|f_0(t_m)|^2}{\overline{n_0^2(t)}} = \frac{\left| \int_{-\infty}^{\infty} F(\omega) H(\omega) e^{j\omega t_m} d\omega \right|^2}{2\pi \int_{-\infty}^{\infty} S_n(\omega) |H(\omega)|^2 d\omega} \quad 3-4$$

At this point we make use of the Schwarz inequality

$$\left| \int_{-\infty}^{\infty} f_1(x) f_2(x) d(x) \right|^2 \leq \int_{-\infty}^{\infty} |f_1(x)|^2 d(x) \int_{-\infty}^{\infty} |f_2(x)|^2 d(x) \quad 3-5$$

The equality in Equation 3-5 holds if, and only if,

$$f_1(x) = k f_2(x) \quad 3-6$$

where  $k$  is an arbitrary constant.

Now we let the two functions in Equation 3-5 be identified with  $H(\omega)$  and  $F(\omega)e^{j\omega t_m}/[S_n(\omega)]^{1/2}$ , respectively, so that Equation 3-5 becomes

$$\left| \int_{-\infty}^{\infty} H(\omega) F(\omega) e^{j\omega t_m} d\omega \right|^2 \leq \int_{-\infty}^{\infty} |H(\omega)|^2 d\omega \int_{-\infty}^{\infty} |F(\omega)|^2 d\omega \quad 3-7$$

Substitution of this result into Equation 3-4 gives

$$\frac{|f_0(t_m)|^2}{n_0^2(t)} \leq \frac{\int_{-\infty}^{\infty} |F(\omega)|^2 d\omega \int_{-\infty}^{\infty} |H(\omega)|^2 d\omega}{2\pi \int_{-\infty}^{\infty} S_n(\omega) |H(\omega)|^2 d\omega} \quad 3-8$$

For the special case in which case the noise is white,  $S_n(\omega) = h/2$ , and we have

$$\frac{|f_0(t_m)|^2}{n_0^2(t)} \leq \frac{1}{\pi\eta} \int_{-\infty}^{\infty} |F(\omega)|^2 d\omega = \frac{E}{\eta/2} \quad 3-9$$

where  $E$  is the energy in  $f(t)$  for a 1-ohm load. The equality in Equation 3-9 holds only if

$$H_m(\omega) = kF^*(\omega)e^{-j\omega t_m}, \quad 3-10$$

or its inverse Fourier transform

$$h_m(t) = \mathcal{F}^{-1}\{kF^*(\omega)e^{-j\omega t_m}\} = kf^*(t_m - t) \quad 3-11$$

The constant  $k$  is arbitrary and we assume  $K=1$  for convenience. And for the

$$h_m(t) = f(t_m - t)$$

We conclude from this result that the impulse response of the optimum system is the mirror image of the desired input signal  $f(t)$ , delayed by an interval  $t_m$ . Hence the filter is matched to a particular signal, as conveyed by the terminology "Matched Filter".

$$g(t) = f(t) * f(t_m - t)$$

The result expressed in Equation 3-10 makes good sense intuitively when applied to the magnitude characteristic of a filter so that  $|\mathcal{H}(\omega)| = |\mathcal{F}(\omega)|$ . This



result states that one should filter in such a way as to attenuate strongly those frequency components in frequency intervals having little relative signal energy while attenuating very little those components where the relative signal energy is high [12]. Recall also that we are filtering for signal recognition in the presence of noise, not for signal fidelity.

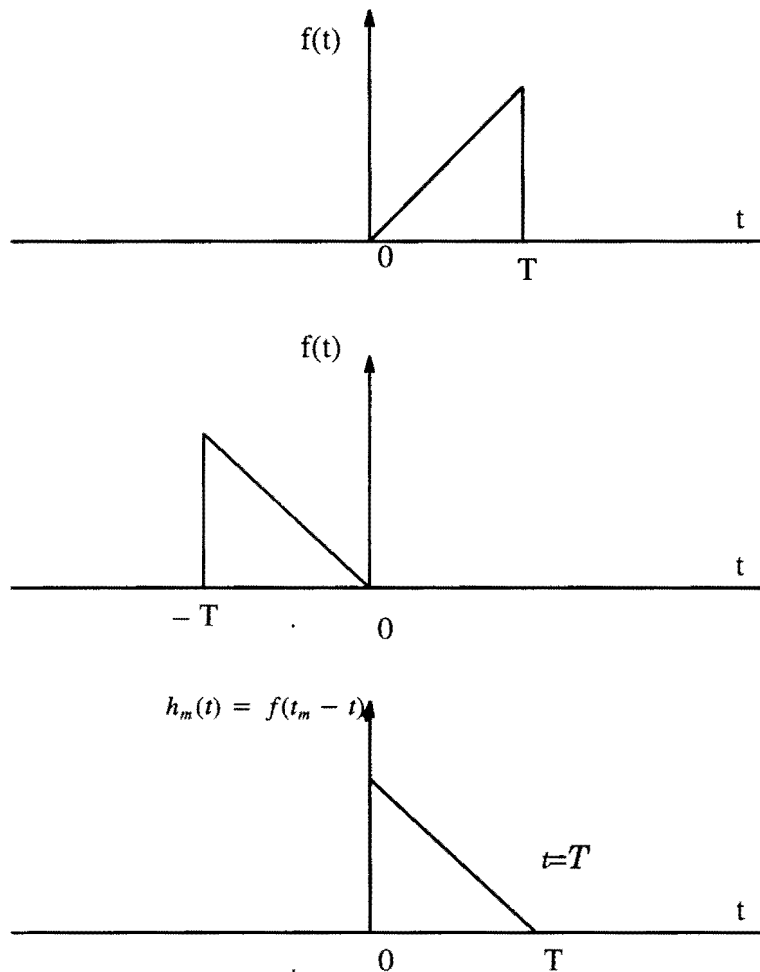


Figure 3.1 The Matched Filter for a Physically Realizable System, minimum delay

The phase response is also very important and Equation 3–10 states that the phase shifts in  $f(t)$  should be negated in such a way that all frequency components in  $f(t)$  add in phase at exactly the time  $t=t_m$ . In contrast, the noise spectral components add with random phases so that the peak–signal–to–rms–noise ratio is maximized.

The signal representation  $f(t)$  is assumed to have a finite duration  $(0,T)$ . The impulse response of the matched filter  $f(t_m-t)$  can be obtained by folding  $f(t)$  about the vertical axis and shifting it to the right by  $t_m$  seconds. Restricting consideration to the physically realizable case with minimum delay, we choose  $t_m=T$ . This is illustrated in Figure 3.1

At the point  $t=t_m$ , the signal output of the matched filter is given by substituting Equation 3–10 in Equation 3–2 with  $\kappa=1$ :

$$f_0 = \frac{1}{2\pi} \int_{-\infty}^{\infty} |F(\omega)|^2 d\omega = E \quad 3-12$$

Thus the output of the matched filter at  $t=t_m$  is independent of the particular waveform chosen and depends only on its energy. The mean–square noise output of the matched filter [Equation 3–12 in Equation 3–9] is:

$$\overline{n_0^2(t)} = E \frac{\eta}{2} \quad 3-13$$

### 3.2.2 Test Algorithm

According to the theory of matched filter, it is a very important step to find the reference signal and then build the matched filter which matches the pre-

dicted nature of the signals that we are going to detect. The algorithm extracts the time-varying spectral contents of the reference signal, and uses the information to filter out the noise outside the relevant band for each component of the signals.

The frequency response of the filter is designed to match content for consecutive time segments of the reference signal. Then, each segment of the signal trial is filtered by the reference filter. The procedure is applied to a series of signal trials, after which the processed trials are analyzed to obtain the clinically-relevant information [13].

1. The reference signal is obtained by ensemble averaging of  $N$  single trials

$$\bar{s}(n) = \frac{1}{N} \sum_{i=1}^N s_i(n) \quad 3-14$$

where  $S_i(n)$  represents each single trial.

2. Find the coefficients of a polynomial  $P_{(x)}$  of degree  $N$  that fits the reference signal

$$P_{(x)} = a_0 + a_1x + a_2x^2 + \dots + a_nx^n \quad 3-15$$

The experimental result showed that the 6th degree  $P_{(x)}$  carries the most important characteristics of the reference signal.

3. Using those seven coefficients, an approximate reference signal is evaluated as a vector  $X_{(n)}$ .

4. Fold  $X_{(n)}$  about the vertical axis and shift it to the right by the duration length of this reference signal ( $L=256$  points). This is the waveform that is going to be used as matched filter  $h_{(n)} = X_{(L-n)}$ .

5. As we know from section 3.2., the convolution integral  $Y_{(n)}$  of  $X_{(n)}$  and  $h_{(n)}$  has the maximum result at time  $t=T$  or  $n=L$ . It is the same result of the auto-correlation of  $X_{(n)}$ . So

$$\begin{aligned}
 Y_{max(n)} &= Y(L) \\
 &= x(n) * h(n) \\
 &= \int_{-\infty}^{+\infty} x(\tau)h(L - \tau)d\tau \\
 &= \int_{-\infty}^{+\infty} x(\tau)x(L - \tau)d\tau
 \end{aligned}$$

and  $\frac{1}{Y_{(L)}}$  will be used as the normalization coefficient.

### 3.3 MATCHED FILTERING OF ERPs

#### 3.3.1 Design of Matched Filter

In clinics, doctors have run thousands of EEG files, during the last two decades, to obtain averaged ERP waveforms. Their experiments report some representative waveforms for normal and memory – impaired subjects. However, those waveforms are not available in electronic format. In our limited data of EEG signals, we found that the averaged ERP waveform of one of the young and normal subjects is closest to the previous average for normal subjects in clinics, according to the knowledge of the doctors. The first reference signal to design the matched filter in this thesis is obtained by averaging 40 single ERP trials of a young and normal subject, shown as Figure 3.2 (a).

After carefully considering energy distribution of the averaged EEG waveform and comparing its approximate polynomial functions at different degrees,

we choose a polynomial  $P_{(x)}$  of degree 6 to approximately fit the waveform in Figure 3.2 (a). This is shown in Figure 3.2 (b). Seven coefficients of the 6th degree polynomial function are calculated.

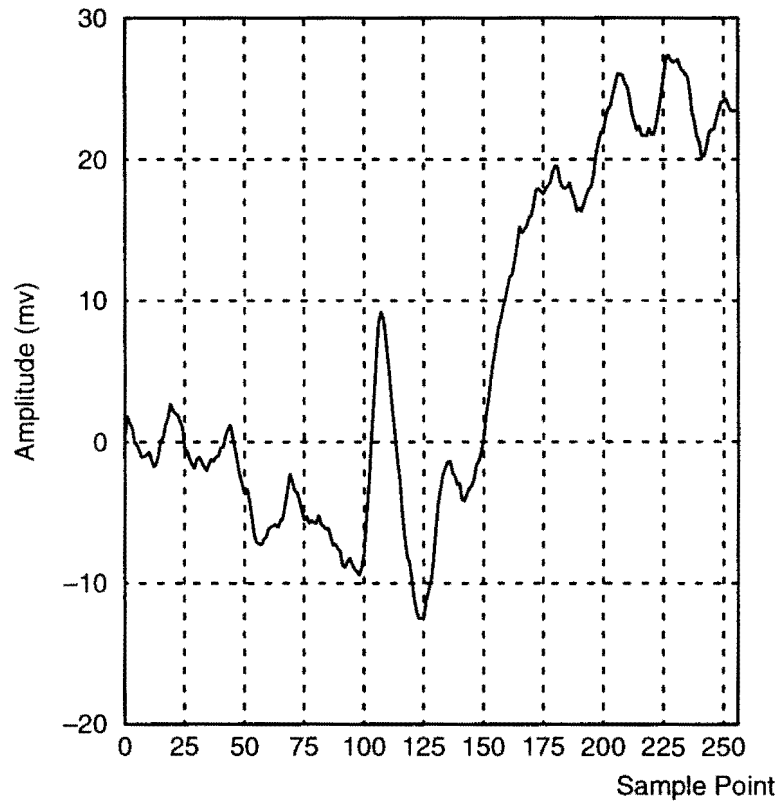


Figure 3.2 (a)Normal AVG File

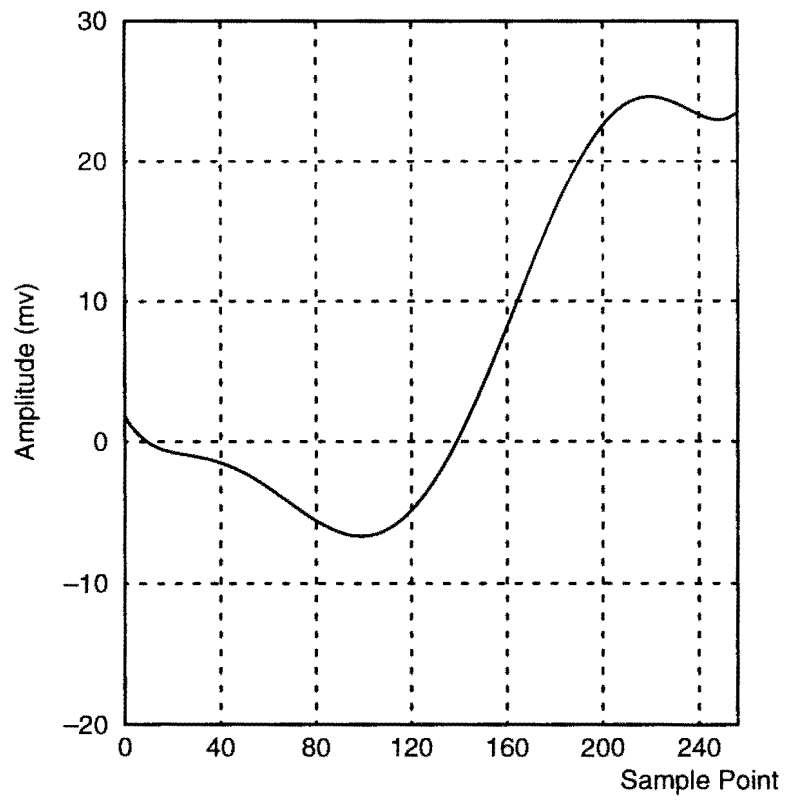


Figure 3.2 (b)  $P(x)$  of Figure 3.2 (a)

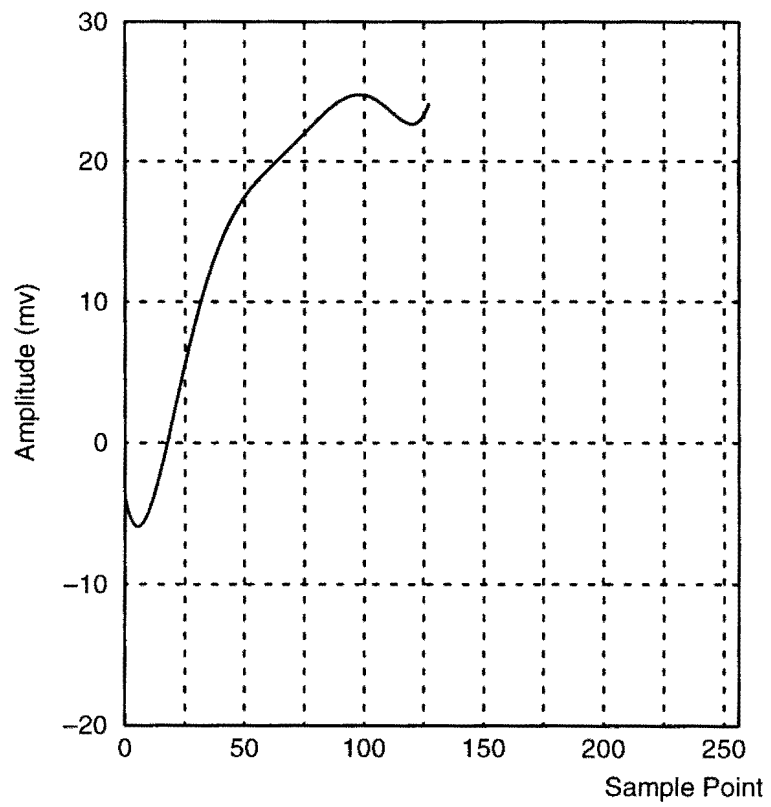


Figure 3.2 (c) The second half of Figure 3.2 (b)

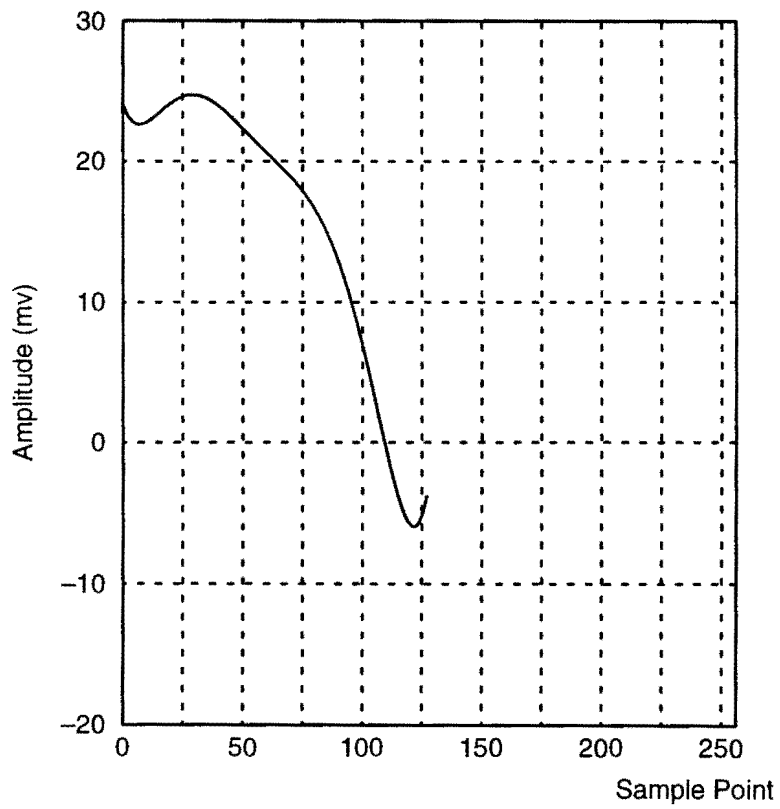


Figure 3.2 (d): The Matched Filter



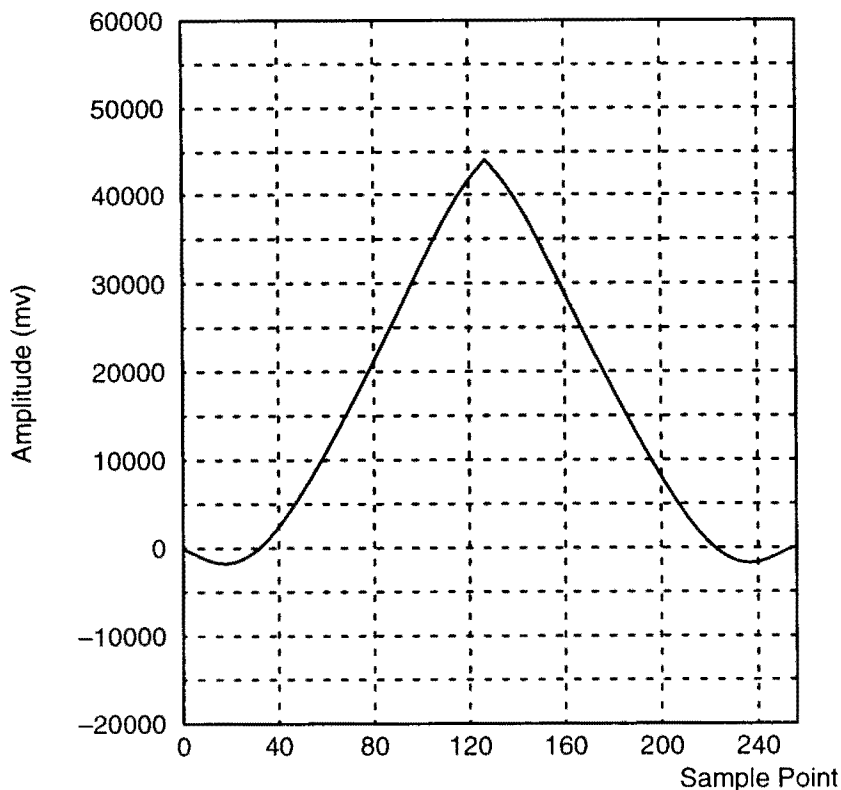


Figure 3.2 (e): Auto-correlation of the matched filter

Figure 3.2 The design procedure of the matched filter

$$a_{0-7} = [0.0000, -0.0000, 0.0000, -0.0031, 0.1121, -1.0840, -2.8531]$$

Using these seven coefficients, the 6th degree polynomial function  $P_{(n)}$  is evaluated at  $1 \leq n \leq 256$ , which is shown in Figure 3.2 (b). We found that the values of the first half of this waveform are all very low. This means that the first half contains very little energy and information. So only the second half is chosen as the characteristic waveform (shown in Figure 3.2 (c)) which will be used to build the matched filter.

Figure 3.2 (d) is the matched filter, after folding and shifting  $P_{(x)}$  128 points to the right. Figure 3.2 (e) is the auto-correlation of  $P_{(x)}$ . We choose the maximum value  $G$  as the energy normalization constant, where

$$G = R_p(128) = \sum_{-\infty}^{+\infty} |p(n)|^2 = 4.405e + 4 \quad 3-16$$

The matched filter output  $Y_{(n)}$  is obtained by convoluting the matched filter with the signal trial.

### 3.3.2 EEG Signal Processing Using Matched Filter

Figure 3.3 (a) is 20 sweeps of the EEG raw signal of a normal subject and Figure 3.3 (b) is the output of the matched filter. Figure 3.4 (a) is 20 sweeps of EEG signal of a memory-impaired subject and Figure 3.4 (b) is the matched filter output. We can observe that after the matched filtering, the ERP waveforms are very clear. Each strip is for one EEG sweep which contains 256 time points and there is one high peak in almost every sweep. The peaks with values beyond the normalization range  $[-1, 1]$  are considered to be caused by the eye-blinks.

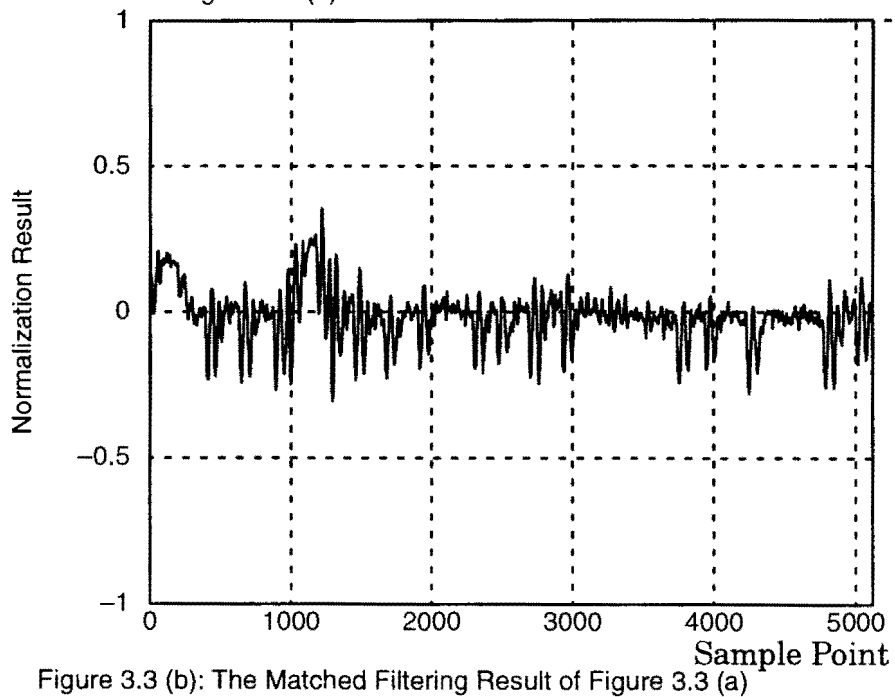
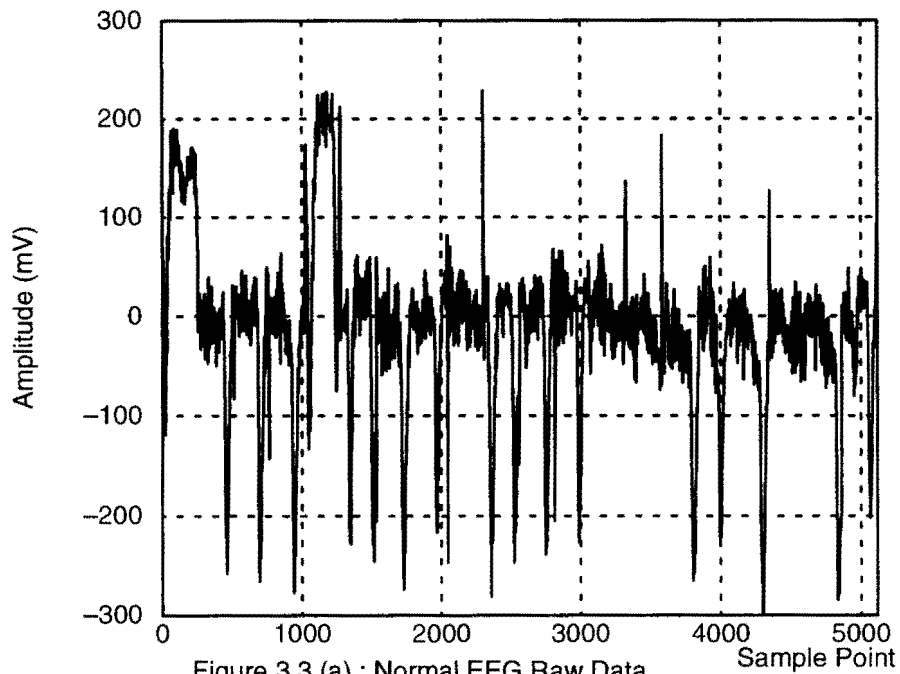


Figure 3.3 The matched filtering result of 20 sweeps of EEG signal of a normal subject

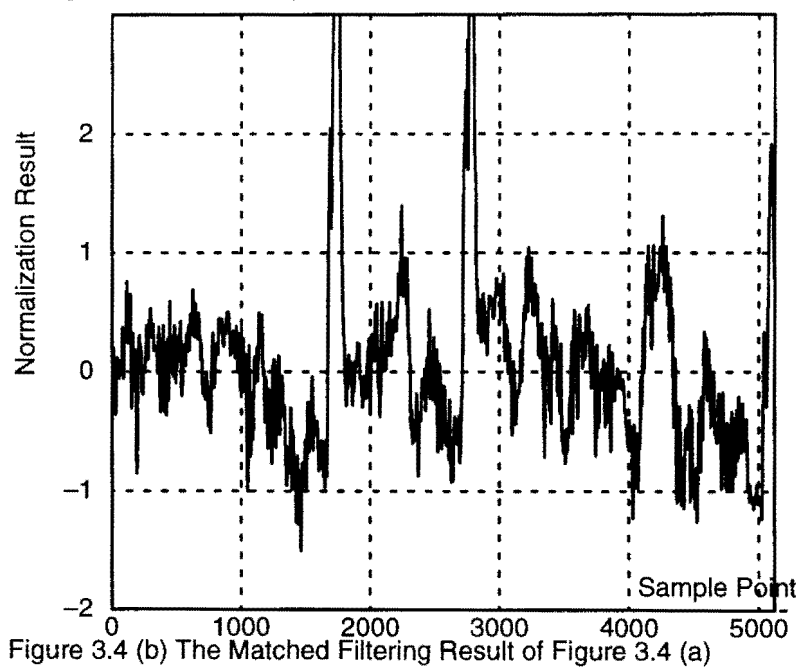
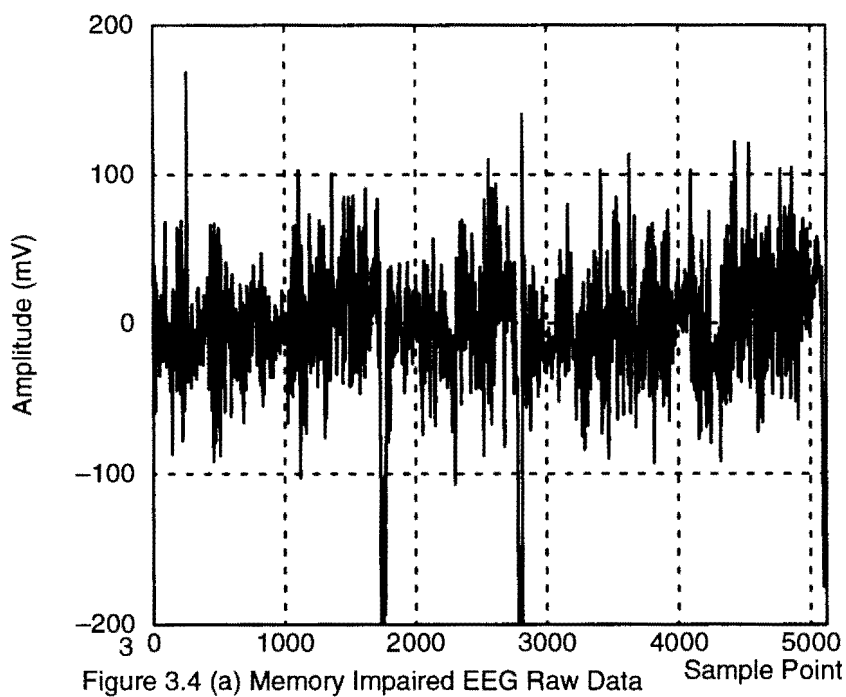


Figure 3.4 The matched filtering result of 20 sweeps of EEG signal of a memory-impaired subject – Page 37

### **3.3.3 ERP Analysis**

Now the ERP analysis can be performed by using the output of the matched filter.

1. Find and record the maximum value in each sweep so there are a total of 80 peak values, corresponding to the 80 sweeps. The time latencies where the peaks occur are also saved in a vector for further analysis.

2. In each set, we are only interested in the target word and its first priming word. While being asked to make a judgement on a relationship or nonrelation of the target word after three priming words, the averaged ERPs suggested that the normal subjects have larger ERP responses than the impaired subjects.

3. We compare the peak values between each target word (both related and unrelated) and the corresponding to its first priming word. The value 0.1, which is the 10 percent of the maximum value of the auto-correlation of the reference signal which was modeled from a normal ERP waveform, is chosen as a threshold. Medically, there is a significant difference between the two conditions. For conditions 2 and 3, the percentages that the peak values are higher than that of their first condition 1 are computed respectively.

4. If either the condition 1 or the target word causes an eyeblink, then the comparison will not be made on that set. The normalization helps to remove the peak values that were caused by the eyeblinks, but the normalization causes some error. We will discuss it later.

### **3.3.4 Analysis Result**

On Table 3.1, Column one is the patient number, totally 12 patients. Column 2 is the patient's group number: Group 1 ( normal subjects group and Group 2 ( memory impaired subjects group). Column three is the patient's recognition percentage by filter MAT12. From Table 3.1, we observe that in the normal subjects group (Group 1), there are only two exceptions, subject 02 and subject 06, whose recognition percentages are below 40%. For all other subjects they are above 40%. As for the impaired subjects (Group 2), the recognition percentages are obviously lower than those of the normal subjects. Only subject 11 is an exception. Its recognition percentage is higher than 30%.

We noticed that for a few cases using the normalization to detect eyeblinks is not very accurate. On one hand, for some memory –impaired subjects, the waveforms are very low, even some eyeblink peaks are still in the range of normalization. And on the other hand, for some very active normal subjects, some regular responses are out of range. Even though those are just a few cases, they did cause some errors in our results.

	Group	MAT12
1	1	48.4
2	1	14.2
3	1	48.4
4	1	50.0
5	1	52.6
6	1	25.0
7	1	75.0
8	1	59.3
9	1	56.7
10	2	12.0
11	2	31.5
12	2	26.3
13	2	16.7
14	2	8.82
15	2	20.0
16	2	5.26

Table 3.1 The ERP Analysis Result Using Matched Filter MAT12

### 3.3.5 Recognition Percentages By Different Matched Filter

In the last section, we used the averaged ERP waveform of a young and normal subject as our reference filter. Now, since we are not quite sure if this EEG file is a very good reference filter, we want to use each EEG file as a reference filter to obtain each **recognition percentage** value. We will use these results in the Mann–Whitney statistics test to get the Matched Filter Bank for our clinic application.

	Group	N1	N2	.N3
1	1	30.3	30.3	9.00
2	1	10.7	10.7	0.00
3	1	36.3	36.3	18.1
4	1	50.0	50.0	50.0
5	1	0.00	0.00	0.00
6	1	15.0	15.0	7.5
7	1	50.0	50.0	25.0
8	1	37.5	37.5	18.7
9	1	54.0	16.2	2.70
10	2	14.8	14.8	0.0
11	2	26.3	26.3	15.7
12	2	5.26	5.26	2.60
13	2	5.55	5.55	2.77
14	2	0.0	0.00	0.0
15	2	2.85	2.85	2.85
16	2	0.0	0.00	0.0

Table 3.2 Recognition Percentage Table for N1 N2 N3



	Gro up	A1	A2	A3
1	1	60.6	39.3	45.4
2	1	32.1	3.57	7.14
3	1	51.5	33.3	39.3
4	1	.00	50.0	50.0
5	1	42.1	36.8	26.3
6	1	17.5	27.5	27.5
7	1	66.6	.00	.00
8	1	62.5	28.0	28.0
9	1	54.0	8.10	8.10
10	2	22.2	7.40	7.40
11	2	42.1	52.6	47.3
12	2	21.0	23.6	18.4
13	2	30.5	13.8	8.30
14	2	16.2	18.9	10.8
15	2	48.5	40.0	25.7
16	2	21.0	10.5	5.26

Table 3.3 Recognition Percentage Table for A1, A2 and A3

	Gro up	JM1	JM2	EX2
1	1	21.0	24.0	21.2
2	1	3.0	3.0	3.57
3	1	24.0	21.0	33.3
4	1	0.0	0.0	50.0
5	1	21.0	21.0	15.7
6	1	10.0	10.0	12.5
7	1	0.0	0.0	25.0
8	1	15.0	15.0	34.37
9	1	5.0	5.0	8.1
10	2	0.0	0.0	11.1
11	2	26.0	26.0	10.5
12	2	0.0	0.0	5.26
13	2	2.0	2.0	2.77
14	2	0.0	0.0	0.0
15	2	2.8	2.0	5.71
16	2	5.0	5.0	5.26

Table 3.4 Recognition Percentage Table for JM1, JM2 and EX2

	Gro up	ES1	ES2	OW1	OW2	AM1
1	1	18.0	21.0	21.2	33.3	21.2
2	1	0.0	3.0	3.57	0.0	0
3	1	21.0	33.0	21.2	12.1	9.0
4	1	50.0	50.0	34.0	50.0	50.0
5	1	0.0	15.0	15.8	26.3	31.5
6	1	12.0	12.0	5.0	7.5	7.50
7	1	41.0	25.0	8.3	25.0	0.0
8	1	31.0	34.0	21.8	25.0	3.12
9	1	8.0	8.0	5.4	5.4	5.40
10	2	3.0	11.0	0.0	0.0	0.0
11	2	10.0	10.0	5.2	15.7	5.26
12	2	2.0	5.0	2.6	0.0	0.0
13	2	5.0	2.0	5.5	5.5	2.77
14	2	0.0	0.0	0.0	0.0	0.0
15	2	2.0	5.0	2.85	2.8	0.0
16	2	0.0	5.0	0.0	5.26	10.52

Table 3.5 Recognition Percentage Table for ES1, ES2, OW1, OW2 and AM1

## CHAPTER 4

# CORRELATION COEFFICIENT AND MANN–WHITNEY TEST

### 4.1 INTRODUCTION

**Correlation coefficient** is a statistical parameter that allows to find the degree of association that exists between two EEG files. A high **correlation coefficient** proves the existence of a close mathematical relationship between the two EEG files. The **correlation coefficient** ranges from  $-1$  to  $+1$ . A minus sign indicates negative correlation, and a plus sign indicates a positive correlation.

If there is a strong association between two EEG files, then knowing the subject memory status of one subject helps in predicting the other's memory status. In the opposite case, the weak association between them makes it difficult to guess the memory status of one subject by knowing the other one. From this point, we calculated the **correlation coefficients** between all the EEG files that were received from the clinic.

The **Mann–Whitney** test is a nonparametric statistical test method. As for the nonparametric method [14], the population, from whose random samples are taken, does not have to be normally distributed. This assumption fits the properties of EEG signals very well. In our research, the Mann–Whitney

test is applied to find whether there exists a significant difference between two population means from Group 1 (the normal memory group) and Group 2 (the memory impaired group).

For the Mann–Whitney test, there are two inputs: one is the **recognition percentages**, the other is the **correlation coefficients**. The significant level for this research was selected to be 0.05. The outputs from the Mann–Whitney test are shown in the later part of this chapter. With **recognition percentage** mean values or **correlation coefficient** mean values, the outputs can distinguish Group 1 from Group 2. Future applications in the clinic application will be discussed in the next chapter.

## 4.2 CORRELATION COEFFICIENT

The statistical techniques that have been developed to measure the amount of association between variables are called the correlation methods. A statistical analysis performed to determine the degree of correlation is called a correlation analysis. The statistics used to measure correlation is the **correlation coefficient**. Therefore, **correlation coefficient** is a measurement of the relationship between two variables.

### 4.2.1 The Mean of Sample

The arithmetic mean, which is simply referred to as “the **mean**”, is the most commonly used average [15]. It is the sum of the values observed divided by the number of observations summed. The statistical average of a random variable  $X$  (or a function of a random variable) is the numerical average of the values which  $X$  (or a function of  $X$ ) can assume, weighted by their probabilities.

If  $X = x_1$  is observed  $n_1$  times,  $X = x_2$  is observed  $n_2$  times, etc., until, finally,  $X = x_k$  observed  $n_k$  times, then  $n_1 + n_2 + \dots + n_k = N$  and the observed value is

$$\bar{X} = \frac{1}{N} \sum_{i=1}^N x_i \quad 4 - 1$$

#### 4.2.2 The Variance and Standard Deviation of Sample

The **variance** of a set of samples give us a method to handle the problem of signs of deviations from the mean. Each deviation from the mean, which is  $|X_i - \bar{X}|$  ( $i=1,2,\dots,n$ ), will be squared, and then the results will be added. By the squaring operation, the deviations from the mean will sum not to 0, but to a positive number. Each deviation will contribute to the sum of squares, regardless of the sign. This sum of squares can be regarded as a measure of the total dispersion of the distribution. By dividing the sum by  $N$ , the number of items in the sample, we obtain the mean of squares of deviations, a measure called the variance of the distribution. As a formula, the variance of a sample set  $X$  of  $n$  observations commonly designated  $s_x^2$ , is

$$s_x^2 = \frac{\sum_{i=1}^n (X_i - \bar{X})^2}{n - 1} \quad 4 - 2$$

The **standard deviation** of a set of samples is the most widely used measure of dispersion. The value of the standard deviation tells how closely the values of a data set are clustered around **the mean**. In general, a lower value of the standard deviation for a data set indicates that the values of that data set are spreading over a relatively smaller range around **the mean**. On the other

hand, a large value of the standard deviation for a data set indicates that the values of that data set are spreading over a relatively larger range around the mean.

The **standard deviation** is obtained by taking the positive square root of the **variance**. The variance calculated for population data is denoted by  $s^2$ . Consequently, the standard deviation calculated for the sample data is denoted by  $s$ . Following are the basic formulas that are used to calculate the variance of a sample set  $X$ .

$$S_x = \sqrt{\frac{\sum_{i=1}^n (X_i - \bar{X})^2}{n}} \quad 4 - 3$$

Because of the operation of squaring, the variance is expressed in the square unit, and not in the original unit. It is therefore necessary to extract the square root to restore the original unit. The measure of dispersion thus obtained is called the standard deviation.

### 4.2.3 Sample Covariance and Correlation Coefficient

For a sample of  $n$  elements with the corresponding pairs of data values  $X_1, X_2, \dots, X_n, Y_1, Y_2, \dots, Y_n$ , similar to the sample variance, the **sample covariance** of samples set  $X$  and  $Y$  is defined by the following equation:

$$S_{xy} = \frac{\sum_{i=1}^n (X_i - \bar{X})(Y_i - \bar{Y})}{n} \quad 4 - 4$$

With definition of sample variance and sample covariance, now we can define the sample **correlation coefficient**:

$$r = \frac{S_{xy}}{S_x S_y} = \frac{\sum_{i=1}^n (X_i - \bar{X})(Y_i - \bar{Y})}{\left[ \sum_{i=1}^n (X_i - \bar{X})^2 \times \sum_{i=1}^n (Y_i - \bar{Y})^2 \right]^{1/2}} \quad 4 - 5$$

$$-1 \leq r \leq 1$$

The **correlation coefficient**  $r$  serves as a measure of the extent to which X and Y are dependent. When  $r=0$ , the random variables X and Y are said to be uncorrelated .

From Equations 4-4 and 4-5, we conclude that if two random variables are statistically independent, then they are uncorrelated. However, the converse is not necessarily true.

The **correlation coefficient**  $r$  should exhibit two characteristics:

1. It should be large when the variables are closely associated and small when there is little association.
2. It must be independent of the units used to measure the variables.

Summary of **correlation coefficient**:

The **correlation coefficient** is a pure number, without units. It is not affected by:



- 1 interchanging the two variables;
- 2 adding the same number to all the values of one variables;
- 3 multiplying all the values of one variable by the same positive number.

### 4.3 CORRELATION COEFFICIENT OF AVG FILES

Considering that we will use statistical methods to analyze the **correlation coefficient**, we calculated all the **correlation coefficient** values of our average files. In the later part of this chapter, the Mann–Whitney test will be applied for these **correlation coefficients**.

We selected 29 independent EEG files to calculate their **correlation coefficients**. On Table 4.1, Column 1 is the patient number, totally 16 patient. Column 2 is the group number of the patient. Column 3 is the **correlation coefficients** between MAT12 and sixteen patients. And so are Column 4 by N1 Column 5 by N2 and Column 6 by N3.

In addition, we calculated the **correlation coefficients** between A1, A2, ..., ES2 and sixteen patients. The results are shown from Table 4.2 to Table 4.8. The total number of coefficient column from Table 4.1 to Table 4.8 is 33.

	Group	MAT12	N1	N2	N3
1	1	-.220	0.00	.000	-.10
2	2	.08	.020	.020	.200
3	1	.3	.360	.360	.390
4	2	.98	.829	.800	.830
5	1	.58	.290	.290	.330
6	2	.21	.310	.310	.230
7	1	.21	-.06	-.06	-.03
8	1	.80	.380	.380	.440
9	1	.31	-.45	.210	.220
10	1	.19	.270	.270	.266
11	1	-.12	-.05	-.05	-.20
12	2	.29	.390	.390	.460
13	2	.27	-.12	-.12	-.11
14	2	-.14	.300	.300	.310
15	1	.13	.020	.000	.000
16	2	.46	.290	.290	.350

Table 4.1 Correlation Coefficient Table for MAT12, N1, N2 and N3

	Group	A1	A2	A3	JM1	JM2
1	1	.00	.460	.370	.23	.18
2	2	-.39	.340	.00	-.14	-.15
3	1	-.28	.310	.00	.0	.0
4	2	.480	-.99	-.98	.99	.90
5	1	.510	.470	.490	.63	.60
6	2	-.14	.070	.00	-.06	.00
7	1	-.20	.300	.360	.31	.50
8	1	.200	.500	.00	-.27	-.20
9	1	-.45	.00	-.16	-.20	-.20
10	1	-.19	.200	.190	.0	.04
11	1	-.36	.240	-.06	-.13	-.13
12	2	-.08	.330	.310	.25	.22
13	2	-.20	-.13	-.30	-.20	-.25
14	2	.35	.360	.260	.23	.09
15	1	-.29	.210	.240	.10	.02
16	2	.250	.480	.360	.30	.36

Table 4.2 Correlation Coefficient Table for A1, A2, A3, JM1 and JM2

	Group	OW1	OW2	JS1	JS2
1	1	-.10	-.11	.33	.28
2	2	-.19	-.02	-.67	-.73
3	1	.00	-.17	.86	.86
4	2	.32	.29	-.05	.05
5	1	-.06	-.11	.84	.85
6	2	.27	-.02	-.70	-.64
7	1	-.04	-.08	1.0	.93
8	1	.24	-.09	.40	.47
9	1	.01	-.24	.60	.68
10	1	.40	.01	.33	.45
11	1	.17	.00	.70	.73
12	2	.03	.08	-.71	-.77
13	2	.23	-.04	.07	.13
14	2	1	.52	-.04	.03
15	1	.00	.35	.26	.22
16	2	-.01	.23	.39	.33

Table 4.3 Correlation Coefficient Table for OW1, OW2, JS1 and JS2

	Group	AM1	AM2	EX1	EX2
1	1	.38	1	-.15	.15
2	2	.42	.00	-.26	-.31
3	1	.00	.35	-.01	.05
4	2	-.26	-.15	1.0	.37
5	1	-.01	.39	.10	.25
6	2	-.36	-.33	.32	-.26
7	1	-.01	.33	-.05	.19
8	1	-.53	.17	.29	.00
9	1	-.40	.25	.17	-.07
10	1	-.57	.07	.47	.03
11	1	-.26	.28	.33	.25
12	2	.08	-.16	-.07	-.29
13	2	-.19	.05	-.01	-.45
14	2	-.18	-.10	.32	.12
15	1	-.02	-.13	.20	.51
16	2	.34	.01	-.05	.44

Table 4.4 Correlation Coefficient Table for AM1, AM2, EX1 and EX2

	Group	ML1	HR1	HR2	RG1	RG2
1	1	.05	.39	.37	-.13	.12
2	2	.01	-.59	-.60	-.20	-.51
3	1	.18	.87	.79	-.06	.46
4	2	-.01	.10	.03	.20	.29
5	1	.04	1	.82	.18	.41
6	2	.20	-.62	-.68	-.35	.19
7	1	.07	.84	.84	.26	.31
8	1	.43	.57	.39	.04	.77
9	1	.47	.69	.55	-.07	.76
10	1	.36	.50	.38	-.11	.78
11	1	.23	.83	.70	.18	.64
12	2	.13	-.64	-.61	-.07	-.23
13	2	1	.04	-.09	-.28	.49
14	2	.23	-.06	-.06	.00	.27
15	1	-.28	.18	.26	1.0	-.24
16	2	-.43	.20	.34	.30	-.37

Table 4.5 Correlation Coefficient Table for ML1, HR1, HR2, RG1 and RG2

	Group	MB2	LM2	SB2	OW1	OW2
1	1	.28	.17	.01	-.10	-.11
2	2	-.62	-.49	-.30	-.19	-.02
3	1	.75	.50	.18	.00	-.17
4	2	.33	.29	-.05	.32	.29
5	1	.83	.57	.20	-.06	-.11
6	2	-.33	.04	-.52	.27	-.02
7	1	.70	.40	.39	-.04	-.08
8	1	.79	1.0	-.51	.24	-.09
9	1	.81	.85	-.28	.01	-.24
10	1	.78	.86	-.37	.40	.01
11	1	1	.80	-.05	.17	.00
12	2	-.48	-.15	-.54	.03	.08
13	2	.23	.43	-.43	.23	-.04
14	2	.17	.24	-.01	1.0	.52
15	1	.18	.04	.30	.00	.35
16	2	-.05	-.51	1.0	-.01	.23

Table 4.6 Correlation Coefficient Table for MB2, LM2, SB2, OW1 and OW2

	Group	MC1	MC2	LS1	LS2
1	1	-.16	-.08	.24	.25
2	2	.73	.68	-.59	-.63
3	1	-.64	-.53	.62	.71
4	2	-.07	-.30	.35	.17
5	1	-.64	-.66	.69	.69
6	2	.54	.23	-.19	-.18
7	1	-.70	-.57	.50	.60
8	1	-.15	-.85	.78	.85
9	1	-.45	-.74	.82	1.0
10	1	-.31	-.73	.80	.78
11	1	-.48	-.83	.80	.81
12	2	1.0	.36	-.49	-.45
13	2	.13	-.12	.25	.47
14	2	.03	-.19	.09	.01
15	1	-.07	-.39	.00	-.07
16	2	-.54	.12	-.19	-.28

Table 4.7 Correlation Coefficient Table for MC1, MC, LS1 and LS2



	Group	CM1	CM2	ES1	ES2
1	1	.00	.13	.10	.35
2	2	1.0	.86	.44	-.61
3	1	-.61	-.43	.08	1.0
4	2	-.26	-.30	-.32	-.01
5	1	-.59	-.42	.04	.87
6	2	.24	.14	-.43	-.58
7	1	-.67	-.53	.09	.86
8	1	-.49	-.45	-.43	.50
9	1	-.63	-.55	-.39	.71
10	1	-.59	-.50	-.52	.50
11	1	-.62	-.51	-.18	.75
12	2	.73	.58	.23	-.64
13	2	.01	-.05	-.12	.18
14	2	-.19	-.09	-.16	.00
15	1	-.20	-.37	.19	-.06
16	2	-.30	-.18	.27	.46

Table 4.8 Correlation Coefficient Table for CM1, CM2, ES1 and ES2

#### 4.4 MANN-WHITNEY TEST

##### 4.4.1 Testing of Hypothesis--Statistical Inference

###### Statistical Hypothesis

A statistical hypothesis is a tentative statement about one or more parameters of a population or a group of populations. We make only tentative statements concerning the parameters, or the state of nature, since we are not per-

fectly certain about the values of parameters. The primary function of inferential statistics is to assist us in reaching sound decisions in spite of uncertainties. As a matter of fact, this function is of such importance that modern statistics has been referred to as the “study of decision making in the face of uncertainty.”

### **Hypothesis Testing**

To test a hypothesis in statistics, sample data are collected and used to calculate a test statistic. The symbol  $H_0$  designates the null hypothesis, and  $H_1$  designates the alternative hypothesis. In most cases, the null hypothesis is the one that asserts the absence of any effect claimed for a certain action or treatment. Depending on the value of the statistic, the null hypothesis  $H_0$  is accepted or rejected. The critical region for  $H_0$  is defined as the range of values of the test statistic that corresponds to a rejection of the hypothesis at some fixed probability of committing a type I error. A type I error means erroneously rejecting the null hypothesis. The test statistic itself is determined by the specific probability distribution sampled, and by the parameters selected for testing [16].

For example, we may test the hypothesis that the population mean  $m$  is equal to  $m_0$  against the alternative hypothesis that  $m$  is not to  $m_0$ ; that is,

$$H_0: m = m_0$$

$$H_1: m \neq m_0$$

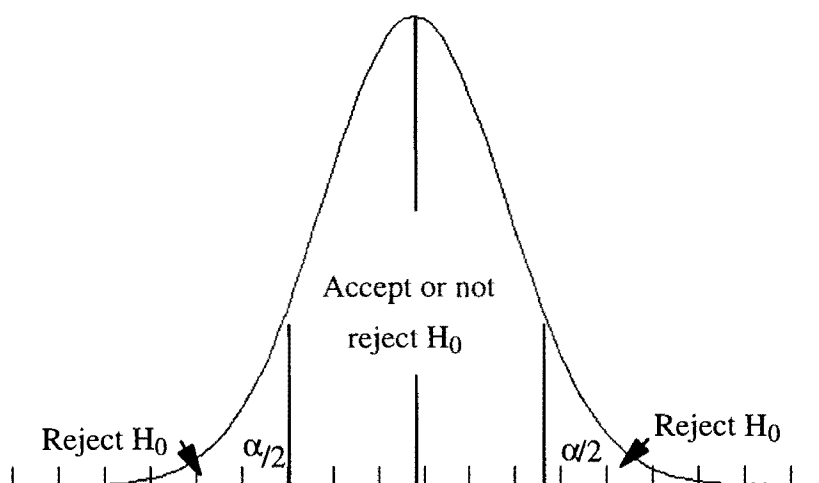


Figure 4.1 Region of rejection for testing  $H_0: m = m_0$  against  $H_1: m \neq m_0$

### Significant Level

The level of significance refers to the state of being “statistically significant”. Once the level of significance is chosen the region of rejection  $\alpha$ , also called the critical region, is decided upon. See Figure 4.1.

### P Value

P values report the smallest level at which the observations are significant, the level of just significance or the critical value. If the P value is smaller than the nominal level, the observations are significant, and otherwise not significant [17].

The general procedure used for testing a hypothesis is as follows:

1. Assume an appropriate probability model to describe the behavior of the random variable under investigation. This choice should be based on previous experience or intuition.
2. Formulate a null hypothesis and an alternative hypothesis. This must be done carefully to permit meaningful conclusions to be drawn from the test.
3. Specify the test statistic.
4. Choose a level of significance  $\alpha$  for the test.
5. Determine the distribution of the test statistic and the critical region for the test statistic.
6. Calculate the value of the test statistic from a random sample of data.
7. Accept or reject  $H_0$  by comparing the calculated value of the test statistic with the values defining the critical region.

#### **4.4.2 Nonparametric Test**

This nonparametric method of statistical procedures is one that does not require knowledge of the form of the probability distribution from which the measurements come. Since nonparametric methods do not require assumptions about the form of the population probability distribution, they are often referred to as distribution-free methods.

From this discussion we see that one reason for using nonparametric methods is that in some situations there is insufficient knowledge about the form of the population distribution. Thus the assumptions, which are necessary for the use of parametric tests, cannot be made.

The second reason for using nonparametric methods concerns the data measurement. The nonparametric methods are often applied to the rank ordered or preference data. Such data differ from the continuous data that we are more familiar with in the sense that the usual numerical measures (e.g., mean, standard deviation, etc.) are not applicable. Preference data are the type of data generated when people express preference for one product over another, one service over another, etc.. Parametric procedures cannot be applied with these data, but nonparametric ones can.

The nonparametric methods, nonetheless, have their advantages by being easy to apply. They are relatively simple and easy to explain and understand [18].

The test, like other nonparametric tests, does not require rigid assumptions about the populations from which samples are taken. The only assumption needed is that the values of the random variable on which two groups are to be compared are continuously distributed. In actual practice, however, no serious difficulty will be encountered even if this assumption is not met.

In this section we present a nonparametric statistical test to determine if there are any differences between the two populations. The nonparametric test is based upon independent random samples from each population.

#### **4.4.3 Mann–Whitney Test Theory**

##### **Test Hypothesis**

The properties of the rank sum will be developed here under the assumption that  $X_1, X_2, \dots, X_{n_1}, Y_1, Y_2, \dots, Y_{n_2}$  are independent observations drawn from

two populations and for the null hypothesis that the populations are equivalent.

The hypotheses tested were

$H_0: m = m_0$  The two populations are identical

$H_1: m \neq m_0$  The two populations are not identical

The null hypothesis to be tested is that two samples independently taken come from two populations having the same mean. Thus the Mann–Whitney test is another useful alternative to the parametric Student T–test when we wish to avoid the assumptions required under the Student T–test. It is also referred to as the U test since the test statistic U is computed from sample data for testing the null hypothesis.

### **U Value with its Mean and Variance**

The U test is usually employed when two independent samples are involved [19]. Suppose that two samples, 1 and 2, with  $n_1$  and  $n_2$  observations respectively, are independently selected and the  $n_1 + n_2$  scores from both samples are arranged in an array of a descending or ascending order [15]. A rank is assigned to each score according to its magnitude. That is, the lowest score is assigned rank 1, the next lowest rank 2, and so on. Then one of the two samples, say sample 1 which has  $n_1$  observations, is chosen, and the sum of its ranks is computed. Let us call this sum  $R_1$ . The test statistic U is defined as

$$U = (\text{Largest Possible Value of } R_1 \text{ or } R_2) - R$$

$$U = n_1 n_2 + \frac{n_1(n_1 + 1)}{2} - R_1 \quad 4-6$$

Similarly, U can be obtained by using the formula

$$U = R_2 - \frac{n_2(n_2 + 1)}{2} \quad 4-7$$

where  $R_2$  is the sum of the ranks in sample 2 with  $n_2$  observations. The two formulas may yield two different values for U. What will be actually used is the smaller of the U values. The larger value is designated U'. Before employing the test, we should check whether we have found U or U' by comparing it with  $n_1 \times n_2 / 2$ . If the resulting value is larger than  $n_1 \times n_2 / 2$ , it is then U', and the value U can be obtained by applying the equation:

$$U = n_1 n_2 - U' \quad 4-8$$

It can be demonstrated that if the two populations are identical, the sampling distribution of U can be approximated for large  $n_1$  and  $n_2$  by a normal distribution with

$$\mu_U = \frac{n_1 n_2}{2} \quad 4-9$$

and standard deviation

$$\sigma_U = \sqrt{\frac{n_1 n_2 (n_1 + n_2 + 1)}{12}} \quad 4-10$$

Accordingly, we determine the significance of an observed U value by computing the standardized normal score

$$Z = \frac{U - \mu_U}{\sigma_U} \quad 4-11$$

#### 4.4.4 Mann–Whitney Test Sample

##### Sample Test Hypothesis

As to illustrate how the U test is used, let us suppose that we wish to determine whether the annual sales volume achieved by salesmen who have graduated from college differs from the volume achieved by have failed to obtain a college degree. The null hypothesis is there is no difference between the two means. Let G and F stand for two groups of salesman, respectively. Suppose further that a random sample of 10 college–graduated salesman ( $n_1=10$ ) and another sample of 21 salesman without a college degree ( $n_1=21$ ) are independently selected. The sales volumes and the ranks are shown in Table 4.9 . Note that, for this example,  $R_1 = 98$  and  $R_2 = 398$ . The value of U is found by substituting the observed quantities in Equation 6 as follows:



Salesman G	Annual sales volume \$	Salesmen F	Annual sales volume in \$
1	82 (in thousands)	1	92
2	75	2	90
3	70	3	90
4	65	4	89
5	60	5	86
6	58	6	85
7	50	7	83
8	50	8	81
9	46	9	81
10	42	10	78
		11	76
		12	73
		13	72
		14	71
		15	68
		16	67
		17	66
		18	64
		19	63
		20	52
		21	40

Table 4.9 Annual sales volumes of college–graduated salesmen, G, and salesmen without a college degree, F.

Sales Volume, \$	Salesmen G	Salesmen F	Rank
40		21	1
42	10		2
46	9		3
50	7		4.5
50	8		4.5
52		20	6
58	6		7
60	5		8
63		19	9
64		18	10
65	4		11
66		17	12
67		16	13
68		15	14
70	3		15
71		14	16
72		13	17
73		12	18
75	2		19
76		11	20
78		10	21
81		9	22.5
81		8	22.5
82	1		24
83		7	25
85		6	26
86		5	27
89		4	28
90		3	29.5
90		2	29.5

92		1	31
	R1=98	R2=398	

Table 4.10 Annual sales volumes and rank of college-graduated salesmen, G, and salesmen without a college degree, F

### Sample U Value with its Mean and Variance

$$U = n_1 n_2 + \frac{n_1(n_1 + 1)}{2} - R_1 = 10(21) + \frac{10(10 + 1)}{2} - 98 = 167$$

which is greater than  $n_1 \times n_2 / 2 = 10(21) / 2 = 105$ . Thus the U value that we should use is

$$U = 10(21) - 167 = 43$$

You should check that Equation 4-7 will yield the same value for U.

$$U = R_2 - \frac{n_2(n_2 + 1)}{2} = 398 - 231 = 167$$

For  $n_1$  and  $n_2$  each less than 20, the smaller value of U is referred to the U table of critical values for determining whether the null hypothesis of no difference between the two means should be rejected. Since in this example  $n_2$  is greater than 20, the normal approximation, and not the U table, will be employed

$$\mu_U = \frac{n_1 n_2}{2} = \frac{10 \times 21}{2} = 105$$

and

$$\sigma_U = \sqrt{\frac{n_1 n_2 (n_1 + n_2 + 1)}{12}} = \sqrt{\frac{10 * 21 * (10 + 21 + 1)}{2}} = 23.66$$

thus

$$Z = \frac{U - \mu_U}{\sigma_U} = \frac{43 - 105}{23.66} = -2.62$$

If the level of significance is chosen at  $\alpha=0.01$ , the critical Z values are  $\pm 2.575$ . Thus we don't reject the null hypothesis, but conclude that the annual sales volume achieved by salesman without a college degree is not equal to the volume for salesman with such a degree.

## 4.5 TEST FOR CORRELATION COEFFICIENTS

### Test Hypothesis

The hypotheses tested were :

$H_0: m = m_0$  The **correlation coefficient** means from two groups are identical.

$H_1: m \neq m_0$  The **correlation coefficient** means from two groups are not identical.

### Test Results for Correlation Coefficient

From previous research, we have all the **correlation coefficients**. Since we eventually want to select the EEG files that could distinguish between the normal memory group and the memory impaired group, the Mann–Whitney test is applied here for MAT12, N1, ..., LS1. The input data is from Table 4.1 to Table 4.8.

After calculating the P value for each averaged EEG file, the test result is as follows in Table 4.11. Column One is the number of averaged AVG file. Column Two is the patient group number. Column Three is the file names we gave. Column 4 is the P values from the Mann–Whitney test for the **correlation coefficients**.

	Group	AVG File	P Value
1	1	MAT12	.223
2	1	N1	.67
3	1	N2	.31
4	1	N3	.56
5	2	A1	.9156
6	2	A2	.83
7	2	A3	.831
8	2	JM1	.71
9	2	JM2	.87
10	1	JS1	.0026
11	1	JS2	.0018
12	1	AM1	.42
13	1	AM2	.0036
14	2	EX1	.71
15	2	EX2	.22
16	2	ML1	.49
17	1	HR1	.0012
18	1	HR2	.0012
19	1	RG1	.31
20	1	RG2	.04
21	1	MB2	.0026
22	1	LM2	.0129
23	2	SB2	.22
24	2	OW1	.31

25	2	OW2	.03
26	2	MC1	.009
27	2	MC2	.0026
28	1	LS1	.005
29	1	LS2	.0036
30	2	CM1	.0005
31	2	CM2	.0036
32	1	ES1	.43
33	1	ES2	.005

Table 4.11 Mann–Whitney Test Results for Correlation Coefficient

## 4.6 TEST FOR RECOGNITION PERCENTAGES

### Test Hypothesis

The hypotheses tested were

$H_0: m = m_0$  The **recognition percentage** means from two groups are identical.

$H_1: m \neq m_0$  The **recognition percentage** means from two groups are not identical.

### Test Result for Recognition Percentage

From previous research, we have all **recognition percentages**. In order to select the EEG files, which could distinguish between the normal memory group and impaired memory group, the Mann–Whitney test is applied here for MAT12, N1, ..., LS1. The input data is from Table 3.1 to Table 3.5.

After calculating the P value for each filter, the test result is as follows in Table 4.12. Column One is the filter number. Column Two is the patient group number. Column Three is the subject names we gave. Column Four the P values from the Mann–Whitney test for the **recognition percentages**.

	Group	Subject	P Value
1	1	MAT12	.0128
2	1	N1	.138
3	1	N2	.0225
4	1	N3	.09
5	2	A1	.112
6	2	A2	.957
7	2	A3	.368
8	2	JM1	.384
9	2	JM2	.196
10	1	ES1	.216
11	1	ES2	.454
12	2	OW1	.0049
13	2	OW2	.0189
14	1	AM1	.0729
15	2	EX2	.0095

Table 4.12 Mann–Whitney Test Results for Recognition Percentage

## CHAPTER 5

### COMPARISONS AND APPLICATIONS

#### 5.1 INTRODUCTION

With the application of the Mann–Whitney statistic test, we obtained the P values for both **recognition percentage** and **correlation coefficient**. In this chapter, we will discuss how to use these test results and when to use them.

Since different statistical tests need very different assumptions, it may be easily confused to use the Student T–test instead of the Mann–Whitney test, which happened in our early research. Here we briefly introduce the Student T–test, its test results and the comparison between the results from these two tests.

In the clinic applications, we set up two banks: The Filter Bank and The Correlation Bank. First, we use every filter in the Filter Bank to calculate the test subject's **recognition percentage**. There is a Mean Value Table for The Filter Bank, which can be used to view the test subject's memory status. Second, we use all the elements of The Correlation Bank to calculate the **correlation coefficients** between the test EEG file and all elements in the Correlation Bank. There is also a Mean Value Table for The Correlation Bank to analyze the linear correlation between the two subjects.



## 5.2 COMPARISONS OF MANN–WHITNEY TEST RESULTS

### 5.2.1 Test Results Analysis

In this research, the significant level is set at  $\alpha = 0.05$  by Dr. Erickson's experience. The Mann–Whitney test results for 22 filters are in Table 4.12. We select the subjects for which the P value is less than 0.05, and build up Table 5.1. These elements can distinguish the **recognition percentage** mean value of normal group from the **recognition percentage** mean value of memory impaired group at significant level  $\alpha = 0.05$ . There are five filters in this table which are AVG files. We name them The Filter Bank. In our application section, we will use each element in this bank to calculate their **recognition percentages**.

On Table 5.1, Column One is the filter number. Column Two is the group number of each filter. Column Three is the the name of each filter. Column Four is the P value of each filter by the Mann–Whitney test.

	Group	Subject	P Value
1	1	MAT12	.0128
2	1	N2	.0225
3	2	OW1	.0049
4	2	OW2	.0189
5	2	EX2	.0095

Table 5.1 The Filter Bank

In the same way, we observe the Mann–Whitney test results for **correlation coefficient** in Table 4.11, select 16 elements, which the P value is less than 0.05 and set up Table 5.2. These elements can distinguish the **correlation coefficient** mean value of normal group from the **correlation coefficient** mean value of memory impaired group at significant level  $\alpha = 0.05$ . It is named The Correlation Bank.

On Table 5.2, Column One is the AVG file number. Column Two is the group number of each AVG file. Column Three is the the name of each AVG file. Column Four is the P value of each AVG file by the Mann–Whitney test.

	Group	Subject	P Value
1	1	JS1	.0026
2	1	JS2	.0018
3	1	AM2	.0036
4	1	HR1	.0012
5	1	HR2	.0012
6	1	RG2	.04
7	1	MB2	.0026
8	1	LM2	.0129
9	2	OW2	.03
10	2	MC1	.009
11	2	MC2	.0026
12	1	LS1	.005
13	1	LS2	.0036
14	2	CM1	.0005
15	2	CM2	.0036
16	1	ES2	.005

Table 5.2 The Correlation Bank

From these two tables, we realize that it needs different subjects for different functionalities. For instance, concerning the **recognition percentage**, it needs MAT12, NORMAL2, OW1, OW2 and EX2 as filters. Concerning the **correlation coefficient**, it needs JS1, JS2, AM2, HR1, HR2, RG2, MB2, LM2, OW2, MC1, MC2, LS1, LS2, CM1, CM2 and ES2 to involve in the correlation operations.

### **5.2.2 Comparison between Two Results**

From Table 5.1 and Table 5.2, we can see that both the matched filter method and the correlation method are efficient in the memory signal processing. They both have produced The Filter Bank and The Correlation Bank which can work on the test subject's EEG file together.

Since there are five filters in Table 5.1, and there are sixteen AVG files on Table 5.2, it is not difficult to tell that the correlation method is better than the matched filter method in this kind of research. However, different methods create different ways to analyze the memory signal, which will be applied in different aspects.

## **5.3 APPLICATIONS OF MANN–WHITNEY TEST RESULTS**

### **5.3.1 Filter Bank**

The Filter Bank is a EEG files set, which includes averaged MAT12, N2, OW1, OW2 and EX2 files. After a test subject raw EEG file is filtered by each

element in the bank, its **recognition percentage** value will be used to adjust whether the subject has normal memory or impaired memory.

### 5.3.2 Filter Bank Means

	FILTERS	Group	RP MEANS FOR NORMAL SUBJECT	RP MEANS FOR IMPAIRED SUBJECT
1	MAT12	1	47.73	22.52
2	N2	1	29.76	7.8
3	OW1	2	15.14	2.3
4	OW2	2	20.51	4.19
5	EX2	2	22.65	5.8

Table 5.3 Mean Table for The Filter Bank

When we analyze the test subject EEG files, the **recognition percentage** after being filtered by the bank element is an important parameter to apply for future research. Table 5.3 is the conference table for the **recognition percentage** means. Column One is the filter number. Column Two is the group number of each filter. Column Three is the mean values of **recognition percentage** for normal subjects by each filter. Column Four is the mean values of **recognition percentage** for memory impaired subjects by each filter. Table 5.4 is almost the same as Table 5.3 except that we create another column to fill up the test subject's **recognition percentages** by each filter so that we can compare the distances between Group One and Group Two

by:

$$D1 = \sqrt{(MAT12N - RPMAT12)^2 + (N2N - RPN2)^2 + \dots + (EX2N - RPEX2)^2}$$

$$D2 = \sqrt{(MAT12I - RPMAT12)^2 + (N2I - RPN2)^2 + \dots + (EX2I - RPEX2)^2}$$

In the above equations and by Table 5.3: MAT12N is 47.73, N2N is 29.76, ..., EX2N is 22.65. MAT12I is 22.52, N2I is 7.8, ..., EX2I is 5.8. RPMAT12 is the **recognition percentage** value of the test subject by the filter MAT12, RPN2 is **recognition percentage** value of the test subject by the filter N2, ..., RPEX2 is the **recognition percentage** value of the test subject by the filter EX2. If D1 is large than D2, the test subject belongs to normal memory group. If D2 is larger than D1, the test subject belongs to memory impaired group.

	FILTERS	GRO UP	RP MEANS FOR NOR- MAL SUB- JECT	RP MEANS FOR TEST SUBJECT	RP MEANS FOR IM- PAIRED SUBJECT
1	MAT12	1	47.73		22.52
2	N2	1	29.76		7.8
3	OW1	2	15.14		2.3
4	OW2	2	20.51		4.19
5	EX2	2	22.65		5.8

Table 5.4 Test Sheet for Recognition Percentage

In clinic application, we first record the raw EEG file of the test subject. Then we calculate **recognition percentages** of this test subject by each filter in the Filter Bank and fill up Table 5.4. The next step is to calculate D1 and D2 to decide which group the test subject belongs to.

### 5.3.3 Correlation Bank

The Correlation Bank is a averaged set of EEG files, which includes JS1, JS2, AM2, HR1, HR2, RG2, MB2, LM2, OW2, MC1, MC2, LS1, LS2, CM1, CM2, ES2. The **correlation coefficients** between the average file of the test subject EEG file and each element in the correlation bank will be supplied to Dr. Erickson to do further memory analysis.

### 5.3.4 Correlation Bank Means

After Table 5.4 is filled up, we calculate D1 and D2. We could decide whether the test subject belongs to the normal memory group or to the memory impaired group by comparing D1 and D2.

Table 5.5 is the conference table for the **correlation coefficient** means. Column One is the average EEG file number. Column Two is the group number of each AVG file. Column Three is the mean values of **correlation coefficient** for normal subjects by each file. Column Four is the mean values of **correlation coefficient** for memory impaired subjects by each file. Table 5.6 is almost the same as Table 5.5 except that we create another column between two mean columns to fill up the **correlation coefficients** of the test subject with each element in the Correlation Bank.

	GROUP	CORRELA-TION SET	CC MEANS FOR NOR-MAL SUB-JECT	CC MEANS FOR IM-PAIRED SUB-JECT
1	1	JS1	.5907	-.2463
2	1	JS2	.6080	-.2280
3	1	AM2	.3009	-.0984
4	1	HR1	.6513	-.2234
5	1	HR2	.5666	-.2387
6	1	RG2	.4453	.0199
7	1	MB2	.6791	-.1080
8	1	LM2	.5764	-.0213
9	2	OW2	-.0493	.1467
10	2	MC1	-.3994	.2594
11	2	MC2	-.5976	.1133
12	1	LS1	.5836	-.1081
13	1	LS2	.6238	-.1263
14	2	CM1	-.4876	.1759
15	2	CM2	-.4034	.1381
16	1	ES2	.6079	-.1704

Table 5.5 Mean Table of The Correlation Bank

In clinic application, we create the averaged EEG file (AVG file) from the raw EEG file of the test subject. Then we calculate **correlation coefficients** of this AVG file with each element in the Correlation Bank and fill up Table 5.6.

First, when the subject in the Correlation Bank represents some special memory characteristic, we predict that this test subject has the same characteristic if the **correlation coefficient** between this file in the Correlation Bank and the test file is near 1. Second, if the **correlation coefficients** between

the files in the Correlation Bank and the test file is near the mean values, the doctor will do further comparisons between both two EEG waveforms and two AVG waveforms to analyze the memory status of the test subject. These applications will be done by the doctor's personal clinic experience.

	GROUP	COR- RELA- TION SET	CC MEAN FOR NORMAL SUBJECT	CC MEAN FOR TEST SUBJECT	CC MEAN FOR IM- PAIRED SUBJECT
1	1	JS1	.5907		-.2463
2	1	JS2	.6080		-.2280
3	1	AM2	.3009		-.0984
4	1	HR1	.6513		-.2234
5	1	HR2	.5666		-.2387
6	1	RG2	.4453		.0199
7	1	MB2	.6791		-.1080
8	1	LM2	.5764		-.0213
9	2	OW2	-.0493		.1467
10	2	MC1	-.3994		.2594
11	2	MC2	-.5976		.1133
12	1	LS1	.5836		-.1081
13	1	LS2	.6238		-.1263
14	2	CM1	-.4876		.1759
15	2	CM2	-.4034		.1381
16	1	ES2	.6079		-.1704

Table 5.6 Test Sheet for The Correlation Bank

## 5.4 COMPARISONS BETWEEN TESTS

### 5.4.1 Comparisons of Two Test Assumptions



Student T–test Assumptions: The population variances must be identical. The population from which random samples are taken must be normally distributed.

Mann–Whitney Test Assumptions: No requirements for both the population variances and the population distribution from which random samples are taken.

The nonparametric tests are used to determine if two populations are identical. The parametric test, such as the Student T–test described before test the equality of two population means [19]. When we reject the hypothesis that the means are equal with the parametric methods, we conclude that the populations differ only in their means. When we reject the hypothesis that the populations are identical using nonparametric tests, we cannot state how they differ. The populations could have different means, different variances, and/or different forms. Nonetheless, if we had assumed that the populations were the same in every way except for the means, a rejection of the null hypothesis using a nonparametric method would have implied that the means differed. The major advantage of the nonparametric methods, however, is that they don't require any assumptions about the form of the probability distribution from which the measurements come [20].

#### **5.4.2 Student T–test**

One of the Student T–test assumption is that the population from which random samples are taken is normally distributed [21]. The normal distribution is one of many probability distributions that a continuous random variable can possess. It is also the most important and most widely used of all the probability distributions. A large number of phenomena, such as the test scores in

a graduate record examination, are normally distributed either exactly or approximately. The continuous random variables representing every different thing in our world have all been observed to have a (approximate) normal distribution.

The Normal Function :

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-(x-\mu)^2/2\sigma^2}. \quad 5-1$$

First of all, it must be assumed that the random variable X is normally distributed .

The Standard Normal Distribution—Z score

$$Z = (X - \mu) / \sigma \quad 5-2$$

For the standardized normal variable Z, Equation 5-1 becomes:

$$f(Z) = \frac{1}{\sqrt{2\pi}} e^{-Z^2/2}. \quad 5-3$$

From Equation 5-2, if X is replaced by its estimate  $\bar{X}$ , and  $\sigma$  is replaced by its estimate  $\hat{s}_{\bar{X}}$

$$t = \frac{\bar{X} - \mu}{\hat{s}_{\bar{X}}} \quad 5-4$$

Then Equation 5-4 is known as the t ratio.

### 5.4.3 Comparison of Two Test Results

On Table 5.7, Column One are the filter numbers. Column Two is the group number of each filter. Column Three is the filter name of each filter. Column Four is the P value

of each filter by the Student–T test. Column Five is the P value of each filter by the Mann–Whitney test.

	Group	Subject	T Test P Value	M–W Test P Value
1	1	MAT12	.0269	.0128
2	1	N2	.0251	.0225
3	2	OW1	.0318	.0049
4	2	OW2	.0219	.0189
5	2	EX2	.0002	.0095
6	1	N1	.0038	
7	1	N3	.0061	
8	1	AM1	.0342	

**Table 5.7** Comparison Table for The Filter Bank

On Table 5.8, Column One are the AVG file numbers. Column Two is the group number of each AVG file. Column Three is the name of each AVG file. Column Four is the P value of each AVG file by the Student–T test. Column Five is the P value of each AVG file by the Mann–Whitney test.

	Group	Subject	T Test P Value	M-W Test P Value
1	1	JS1	.0001	.0026
2	1	JS2	.0001	.0018
3	1	AM2	.0001	.0036
4	1	HR1	.0001	.0012
5	1	HR2	.0001	.0012
6	1	RG2	.026	.04
7	1	MB2	.0001	.0026
8	1	LM2	.0065	.0129
9	2	OW2	.0165	.03
10	2	MC1	.0005	.009
11	2	MC2	.0018	.0026
12	1	LS1	.0008	.005
13	1	LS2	.0003	.0036
14	2	CM1	.0032	.0005
15	2	CM2	.0108	.0036
16	1	ES2	.0001	.005
17	2	A1	.031	
18	2	A2	.0007	
19	2	A3	.0001	
20	1	MAT12	.0283	
21	1	N1	.0001	
22	1	N2	.0001	
23	1	N3	.0001	

Table 5.8 Comparison Table for The Correlation Bank

## 5.5 CONCLUSION

Since EEG signal analysis is very difficult, our research has involved developing several quite different methods. As for the successful ones, first we introduced the matched filter to improve the signal-to-noise ratio (SNR) of ERPs. Removing eyeblinks in the EEG files, before using the matched filter, further improved the poor SNR. **Recognition percentage** of the matched filter gives us one parameter concept about a subject's EEG file.

Second, **correlation coefficient** is another parameter concept, which uses AVG files instead of the original EEG files. It is a statistical analysis method to measure the degree or the amount of association between two EEG files. This parameter is very important as long as we know about one of the subject's memory status, also if the **correlation coefficient** between these two files is high, and then we can deduce about the test subject's memory status.

Third, with the application of the Mann-Whitney statistic test for both **recognition percentages** and **correlation coefficients**, we set up two banks: The Filter Bank and The Correlation Bank. They are very useful in the clinic application since the **recognition percentage** and **correlation coefficient** by each element of the banks can distinguish whether the test subject belongs to Group One ( normal memory group ) or Group Two ( impaired memory group ).

There have been also many unsuccessful methods that we have used in this research for the memory processing like using wavelet transform. With more EEG files in the further clinic application, we believe that our Filter Bank and Correlation Bank will be improved, and provide more potential applications as

we discussed in Section 5.3. Dr. Erickson can discover more useful applications from these test results in the further memory processing research.

With the guidance of Dr. Li and Dr. Erickson in the last three years, I worked at creating **recognition percentages** myself (working together with Xueming Li for three months), creating **correlation coefficients** myself, creating eye–blink data files myself, applying both the Student–T test and the Mann–Whitney test to our research myself and using the Mann–Whitney test results in the clinic applications myself.

## REFERENCES

- [1] Jhon W. Rohrbaugh, R. Parasuraman and Ray Johnson, " Event-related brain potentials ", QP376.5 .E94, pp.384–390, 1990
- [2] Samuel K. Law, John W. Rohrbaugh, Charles M. Adams and Michael J. Eckardt, " Improving Spatial and Temporal Resolution in Evoked EEG Reposes Using Surface Laplacians " Electroencephalogram and Clinical Neurophysiology, Vol. 88, pp. 309–322, 1993
- [3] R. Naatanen, P. Paavilainen, K. Alho and K. Reinikainen, " Do Event-Related Potentials Reveal the Mechanism of Auditory Sensory Memory in the Human Brain ", pp. 217–221, 1989.
- [4] F. Grandori, G. L. Romani and M. Hoke, " Auditory Evoked Potentials and Fields, pp. 157–180, 1993
- [5] M. Scherg, J. Vajsar and T.W. Picton, " A Source Analysis of the Human Auditory Evoked Potentials " pp. 336–355, 1989
- [6] Jr. Vaughan, W. Ritter and R. Simson, " Electrical Potentials, Behavior and Clinical Use ", pp. 279–290, 1980
- [7] Kai-Bor Yu and Clare D. Mc Gillem, " Optimum Filters for Estimating Evoked Potential Waveforms " IEEE Trans. on Biomedical Engineering, Vol. BME-30, No. 11 pp.730–737, November 1983
- [8] Emanuel Donchin, " Cognitive Psychophysiology: Event Related Potentials and the Study of Cognition " BF311 .C55193 , PP. 397–413, June 1984
- [9] Kenneth R. Erickson research proposal, " Event Related Potential Changes in Memory Dysfunction " pp. 6–10, December 1984

- [10] Prem S. Mann, “ Introductory Statistics ” pp. 505–517, 638, 1995
- [11] D. H. Lange, H. Pratt and G. f. Inbar, “ Matched Filtering of Evoked Potentials: A Real Time Approach ” *Electroencephalogram and Clinical Neurophysiology*, Vol.88, pp.187–191, March 1993
- [12] William A. Gardner, “ Introduction to Random Process with Applications to Signals & System ”, pp. 71, 291–294, 1990
- [13] Ferrel G. Stremler, “ Introduction to Communication Systems ”, pp. 432–435 Tk5103.S74, 1990
- [14] Xueming Li, “ ERP Analysis Using Matched Filtering and Wavelet Transform ”, Thesis for the Degree of Master of Science in Electrical Engineering, Portland State University, 1994
- [15] John W. Pratt and Jean D. Gibbons, “ Concepts of Nonparametric Theory ”, Springer Series in Statistics, pp. 17–34, 249–267, 1971
- [16] Jean Dickinson Gibbons, “ Nonparametric Statistical Inference ” pp.140–149, 1971
- [17] David R. Anderson, Dennis J. Sweeney and Thomas A. Williams, “ Introduction To statistics ” An applications Approach pp. 231–240 , 406–410, 1981
- [18] Robert M. Bethea, Benjamin S. Duran and Thomas L. Boullion, “ Statistical Methods for Engineers and Scientists ”, pp. 177–178
- [19] Lincoln L. Chao, “ Statistics Methods and Analyses ” pp. 75–100, 173–194, 221–227, 254–266, 1974
- [20] Lincoln L. Chao, “ Statistics Methods and Analyses ” pp. 436–448, 1974



[21] Prem S. Mann, "Introductory Statistics" pp. 505–517, 638, 1995