This is to certify that the
thesis entitled

Evaluation and Comparison of Data Reduction and Source Separation Techniques For Event Related Potentials

presented by

Jacob Swary

has been accepted towards fulfillment
of the requirements for the

M.S. degree in Electrical & Computer Engineering

Major Professor's Signature

08/24/2007

Date

Evaluation and Comparison of Data Reduction and Source Separation Techniques For Event Related Potentials

By

Jacob Swary

A Thesis

Submitted to
Michigan State University
in partial fulfillment of the requirements
for the degree of

Master's in Science

Electrical and Computer Engineering

2007

ABSTRACT

Evaluation and Comparison of Data Reduction and Source Separation Techniques For Event Related Potentials

By

Jacob Swary

Study of event-related potentials (ERP), which measure the brain response to specific presented stimuli with electroencephalography (EEG), is the focus of this thesis. In the past, averaging of multiple trials has been used to evaluate ERPs. This ignores the trial-to-trial variability of the brain's response, and has only produced the knowledge of certain response peaks and how they are generally related to some tasks. Recently, attempts at extracting the actual underlying sources generated by the brain are being made to effectively evaluate the brain's response. A common assumption is that the underlying sources are statistically independent, and independent component analysis is used in this blind source separation (BSS). To avoid the assumption that sources are independent in BSS, we are proposing to solve the problem with quadratic time-frequency distributions of the data. In this way, the assumption that sources are sparse in the time-frequency plane, i.e. most data points are close to zero, is applied. Due to sparsity, methods have been developed to estimate first, a mixing matrix, which determines the weighting of each source at each electrode, and then the sources. The two stage approach solves for a number of sources greater than the number of electrodes used in the EEG measurement. This two stage approach and ICA are both applied to a set of measured ERPs and the results are compared in this thesis. It is shown that the proposed method is more effective at extracting well localized components in time and frequency than ICA. These components are shown as comparable at representing the original ERP data variance with ICA.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

CHAPTER 1

INTRODUCTION

1.1 Electroencephalography

Electroencephalography (EEG) is the non-invasive process of measuring electrical potentials from activity in the brain. Electrodes with a conducting medium are placed at multiple locations around the scalp at fixed locations. Electrode placements are used according to the international 10-20 setup as defined in [1]. Potentials measured at each electrode are not necessarily due to the activity in the immediate proximity of that electrode because of the volume conduction in the brain and across the scalp [2]. The measured potentials are amplified, filtered, and stored for processing. An example electrode setup, typical EEG reading, and the electrode locations of the 10-20 standard are shown in Figure 1.1.

(a) (b) (c)

Figure 1.1. (a) An electrode array connected to a person. (b) A graph of typical EEG readings where each row in the graph represents the measurements from one electrode. (c) The locations of electrodes in the 10-20 system.

Excitatory postsynaptic potentials in neurons of the brain cause electrical current to flow through their cell membranes through the cell back through the membrane into surrounding fluid, and inhibitory postsynaptic potentials cause current to flow to the fluid first back to the cell. The potential created from this individual cellular current flow cannot be measured at the scalp. However, when regions of neural tissue in the neocortex are aligned in parallel, and there is synaptic activity over an area of approximately 1 cm^2, the potential can be measured at the scalp and can be considered as produced by a single dipole source [3]. The potentials measured from electrodes on the scalp are due to all such regions in the brain active at any point in time.

In the 1920's Hans Berger first measured the EEG in humans. Potential recordings were initially printed out directly onto paper for visual inspection. Visual inspection by expert neurologists taking into account the waves' frequency, amplitude, spatial distribution, continuous or transient patterns has been typically the main form of EEG analysis [4]. In early attempts to provide quantitative analysis, techniques were developed that tried to mimic the analysis of the expert neurologists, but were only marginally successful due to the subjective nature of their analysis [5]. With the advent of the computer, EEG measurements could be recorded in continuous form (analog) on tape, or sampled and stored in discrete form (digital) for further processing. The introduction of the fast-fourier transform (FFT) algorithm, spectral analysis using the Fourier transform (FT) became a plausible analysis technique.

Before EEG, there was no way of quantifying brain activity and EEG quickly garnered much interest. It was thought that these recordings could be used to gain insight to the mental processes of the human mind and to make connections to brain function and human behavior. EEG is used in many different areas, for example, different stages of sleep can be determined [5]. EEG is used in analysis of epilepsy to find epileptic areas and decide a patient's suitability for surgery [6]. To ensure

recordings of seizures, long evaluation periods are required in this analysis, as such the non-invasive, cheap, and portable characteristics of EEG lend it to the evaluation over other neuroimaging techniques. EEG can be used in the objective observation of emotion changes in diagnosis and evaluation of emotive disorders in which the patient may not be able to explain the changes [7]. Recently, Brain-Computer Interfaces (BCI) are being developed with EEG. A BCI is used to convey messages to the surrounding world not through the traditional neural pathways, but from direct electrical signals measured from the EEG. This is a very important development for those who lack muscle control, either by injury or disease. Because EEG is cheap, and has the time resolution to provide rapid enough communication, its use is desired in BCIs. Work has been done so that the BCI user can move an object on a computer screen, or type messages by alphabet elimination, letter by letter. In the future, it is hoped BCIs can be developed for reliable control of wheelchairs or prosthetics [8]. There is still much to be desired in using EEG to unlock the mystery of brain function.

While positron emission tomography (PET) and functional magnetic resonance imaging (fMRI) have become popular more recently, there is still a need for effective EEG analysis. The use of PET and fMRI provide much higher spatial resolution of activity in the brain than is provided by EEG. However, EEG provides temporal resolution on the order of 1 ms while fMRI is on the order of 1 s and PET 1 min [15]. Both EEG and fMRI are non-invasive techniques, but EEG can be measured from subjects remotely.

1.2 Event Related Potentials

To gain insight to the mental process with respect to perception, memory, and attention, among others, Event Related Potentials (ERP) or Evoked Potentials (EP) are often studied. An ERP is the response in the brain specific to a present audio, visual, or electrical stimulus, i.e., the potential measured that is directly related to

a specific event. Measurements of ERPs are taken with EEG since neural response occurs in milliseconds and EEG has the temporal resolution to successfully capture the response. In ERP measurement, a subject is presented with a simple stimulus and response is required, perhaps the push of a button. This is repeated multiple times during the experiment. The EEG is recorded through the experiment and one trial is considered the EEG measurements time-locked around presentation of one stimulus. Trials are repeated with enough time in between presentation of a new stimulus so that the previous ERP response is complete, and the time between stimuli is varied so that the subject does not develop expectation of rhythmic presentation.

ERPs are "regarded as manifestations of specific psychological processes" [9], and therefore are critical in the study of brain function. The main problem in ERP measurement is that the ERP waveform is many times smaller than ongoing processes in the brain, and so the ERP is difficult to see in a raw EEG measurement. For this reason, ERP analysis has been dominated in the past by analysis of multiple trial averages due to the nature of the measurements. Recently much work is being done in finding new analysis techniques of ERPs. This thesis is focused on the field of EEG study pertaining to ERPs. An example of an experimental setup to record ERPs is shown in Figure 1.2.

1.3 Spectral Analysis of EEG

Early study of the EEG made it apparent that frequency response was an important characteristic of the waves. Spectral activity is generally classified by the different frequency bands. The low frequency activity from 1-3 Hz is the delta band, 4-7 Hz is the theta band, 8-13 Hz is alpha, 14-30 Hz is beta, and above 30 Hz is the gamma band. A general relaxed wakeful state consists of strong rhythmic alpha activity, with some beta and with small arhythmic delta and theta activity. Changes in wakefulness, different sleep stages, disease and injury all change the typical EEG activity [10]. The

Figure 1.2. An experimental ERP setup with visual stimulus presented on computer screen and response required by pushing of button.

FFT can be used to analyze different frequency bands, but also autoregressive (AR) modelling of the EEG can show the spectral characteristics of the recorded waves.

The EEG wave can be represented by an AR model as follows:

$$x(n) = w(n) - a(1)x(n-1) - a(2)x(n-2) - \cdots - a(p)x(n-p), \qquad (1.1)$$

where $x(n)$ is the data sequence, $a(n)$ are the AR parameters, p is the AR order, and $w(n)$ is assumed to be white noise with flat power spectrum and is the error in prediction. The z-transform of equation 1.1 is:

$$X(z) = \frac{W(z)}{1 + a(1)z^{-1} + a(2)z^{-2} + \cdots + a(p)z^{-p}}, \qquad (1.2)$$

$W(z)$ is constant. The power spectrum, $P(z)$, is the magnitude squared of equation 1.2. The poles of $X(z)$ are the roots of its denominator:

$$z^p + a(1)z^{p-1} + \cdots + a(p). \tag{1.3}$$

This is evaluated on the unit circle, $z = e^{j\omega}$, and is factored to give:

$$(e^{j\omega} - P_1)(e^{j\omega} - P_2)\cdots(e^{j\omega} - P_p), \tag{1.4}$$

such that P_i, $i = \{1, \ldots, p\}$ are the complex poles of $X(z)$. The real ω for which $e^{j\omega}$ matches the phase of P_i represents the real poles of $X(z)$, and there is a peak in the frequency spectrum of the data at this value of ω. Evaluating equation 1.2 or its square at $e^{j\omega}$ gives signal amplitude and power respectively [11].

Spectral analysis assumes stationarity of a signal, and thus is not always accurate of the underlying frequency contents since the EEG is very non-stationary. If the length of data is short, the stationarity assumption may hold well enough. Due to the transient nature of ERPs, spectral analysis is not preferred since it cannot capture events that are localized in time. Still, it provides quality results in analysis of the ongoing EEG processes, and has specifically been studied with regard to drug use, injury, disease, and sleep. In general, an adult in a relaxed waking state shows dominant alpha activity, and this activity is diminished when the state changes, either toward alertness or drowsiness [10]. It has been shown that minor head injuries slow the frequency of the ongoing EEG in the alpha band, shifting the spectral peak to the left [12]. Spectral analysis has shown an upward shift in overall spectral power, including a decrease in theta and delta bands and an increase in alpha band during alcohol absorption, and the reverse effect in alcohol elimination [13]. During sleep, the EEG waves are very slowly changing, and so spectral analysis is very useful in

identifying different stages of sleep.

1.4 Analysis of ERPs: Multiple Trial Average

Historically, the solution most widely used in the analysis of ERP signals has been that of averaging multiple ERP trials of identical stimulus presentation together. The EEG measurement of an ERP can be desribed as the ERP plus all remaining activity as:

$$r(t) = s(t) + n(t), \tag{1.5}$$

where $r(t)$ is the total measured response of the trial, $s(t)$ is the ERP, and $n(t)$ is all non-ERP activity. It is assumed that the ERP is deterministic and, since each trial is identical, each ERP response will be the same. Also, $n(t)$ is modeled as a stationary random process. Assume N trials are recorded, the ith trial is written as:

$$r_i(t) = s(t) + n_i(t), \tag{1.6}$$

since $s(t)$ is assumed deterministic, it is modeled the same over all trials, while the total response and background activity depend on the trial number, i. The background activity is modeled such that its mean is zero and variance is σ_n^2. By calculating the average of the response over all trials, $\bar{r}(t) = \sum_{i=1}^{N} r_i(t) = Ns(t) + \bar{n}(t)$, the ERP response can be magnified compared to the background activity. The signal-to-noise ratio (SNR) of the estimated ERP becomes proportional to the number of trials averaged, and so the ERP response becomes clearer with the averaging of more trials [5].

Giving equal weight to each trial can result in some waveforms with very different shapes due to muscle or eyeblink artifacts, or large ongoing EEG signals distorting the final average. One attempt used to reject outliers is to compare responses and

determine a weighting scheme so that the sources are estimated as:

$$\hat{s}(t) = \frac{1}{D}\sum_{i=1}^{n} w_i x_i(t), \tag{1.7}$$

and $D = \sum_{i=1}^{n} w_i$. A whitening filter is applied to the recordings to give $\tilde{r}_i(t)$, and covariance coefficients $\gamma_i = \frac{1}{N}\sum_{t=1}^{T} \tilde{r}_i(t)\bar{r}(t)$ are calculated between each filtered recording and the average of all other recordings. The weights are then:

$$w_i = \begin{cases} 0 & \text{for} \quad \gamma_i \leq 0 \\ \gamma_i & \text{for} \quad 0 \leq \gamma_i \\ C & \text{for} \quad \gamma_i > C \end{cases} \tag{1.8}$$

and C is chosen empirically, usually about 0.8 [5].

The problem with the averaging technique is that ERPs are not deterministic; exact repetitions of the experiment will not lead to the same responses at electrodes. From trial to trial there will be latency deviations of peaks, different amplitudes, and different wave shapes. Deviations occur due to many factors of subject's perfomance, such as attention, expectation, arousal, strategy and others [14]. The average does not truly represent an ERP due to these variations across trials and cannot reflect changes in the subject's state.

1.4.1 Woody Average

Considering the fact that latencies of ERPs vary from trial to trial, SNR of the averaged ERP could be improved if the shifts in time can be detected and corrected. The Woody average is a process that attempts latency correction [5].

The process described in section 1.4 is first used and the averaged ERP is used as a starting template. Each of the individual ERP responses is then cross-correlated with the template. These ERP waveforms are then shifted by the amount that their maximum cross-correlation value took place at. The total average is then recalculated and

this becomes the new template. This process is repeated until the cross-correlation does not change significantly between the current iteration and the previous iteration.

This will provide improved waveforms in the cases where components of each ERP signal all shift the same amount. Separate components have latency changes that occur independent of each other. Thus, the Woody average will still blur components of the ERP response. Further, it is possible that the Woody average will align certain background processes that were independent in individual trials.

1.4.2 Latency Corrected Average

An attempt to improve on the Woody average is made with latency corrected average (LCA) [5]. Peaks of different components in an ERP are aligned from trial to trial so that the SNR of these components in the average is increased.

The mean is removed from each trial and the variance is normalized. Peaks are located in each recording such that the magnitude of the slope on either side of the peak exceeds a set minimum. The points of the sample mean are used in comparison to the detected peaks. If the points of the mean and peak are of the same sign and statistically exceed a specified confidence interval that both are from a zero mean population, then the detected peak is considered a component, otherwise it is not. Peaks over all trials that are of the same sign as the sample mean at that point and between the same zero crossings of the sample mean are considered the same component. The segments around the peaks are then used in calculating a new mean. This mean is not a complete waveform since it is computed from discrete components of waveforms, so interpolation must replace the empty time slots.

This is an improvement on taking the straight sample mean of responses as it improves SNR for different ERP components, and can show additional components than shown by just the mean. It still is not ideal. It only makes up for variations in peak latency, nothing else. In addition, variations in the individual responses are still all classified as one response in the end. All changes in subjects' states are still

being ignored and an opportunity to gain more insight to mental processes is lost.

1.5 ERP Time Components

Through averaging in ERP experiments, multiple different components have been identified related to certain tasks. The results from averaging trials under the same conditions are found and peaks are analyzed with respect to scalp distribution, polarity, and latency. Components are usually identified by their polarity and latency. For example P300 stands for a positive peak after 300 ms. The components presented here are described in [9].

The earliest components that have been found occur within the first 100 ms of stimulus presentation. These components change as a function of the stimulus with regard to intesity and frequency, and can also be affected by attention. This activity is automatic and is related to the signal picked up at the sensory site being transmitted to the brain's processing systems.

One common component is called the mismatch negativity (MMN). This is a negative peak occuring around 100-200 ms after stimulus, and is beleived to occur in the auditory cortex. This type of response is found when two different classes of auditory stimuli are presented when the subject's attention is on a separate task, one class is presented more often than the other. The average is taken over all trials of each stimulus type and the average of the rare stimulus trials is subtracted from the average of the frequent stimulus trials, and the resulting component is the MMN. The MMN represents a sort of mismatch detector. It is an automatic, preattentive processing of deviant features. Because it can be recorded with a delay usually only up to 10 seconds between stimuli, the MMN is a transient type of memory. As a result of the two classes of stimuli differing with more than one factor, the MMN has a larger amplitude, and so may reflect a parallel processing of multiple factors.

The N200 component is a negative peak at about 200 ms and found in either the

auditory or visual cortex, depending on the stimulus. This component represents a comparison that is actively generated by the subject. It follows from a mismatch between two stimuli, or between the stimulus and a mentally formed template. It is different from the MMN in that the subject is paying attention to the stimulus, and the N200 does not necessarily mean two stimuli are being presented. It appears when the stimulus mismatches the subject's expectancy, whether it be from previous stimulus or a memory template. Its latency covaries with response time, so it may be that the N200 shows the feature discrimination happening, which influences the response time. The N200 can be seen in Figure 1.3.

A very commonly studied component is the P300, which is a positive peak around 300 ms, and is strong in the posterior scalp locations. With more than 30 years research, there is no indication of the underlying sources contributing to the P300. It is thought to be a summation of activity over multiple generators. Its amplitude is affected by perceived stimulus probability, but only when the stimulus is relevant to the task at hand. The amplitude is also directly proportional to the processing demand of the task. The peak's latency and reaction time increase when the task is accuracy oriented instead of speed oriented. The latency is also longer when a categorization task is more difficult. The P300 may reflect the updating of mental environment models or the context in working memory. The P300 is shown in Figure 1.3.

A similar component is the Frontal P300, which instead of being strong in the posterior electrodes, is strong in the frontal. This is caused by deviant stimuli which are very rare and unexpected, such that there is no memory template beforehand. In young adults, repeated presentations will shift the P300 response to the posterior. In older adults, the Frontal P300 stays frontal with repeated presentations.

Another examined component is the N400, a negative peak around 400 ms which is related to reading tasks. The N400 response is strong when a string of words is

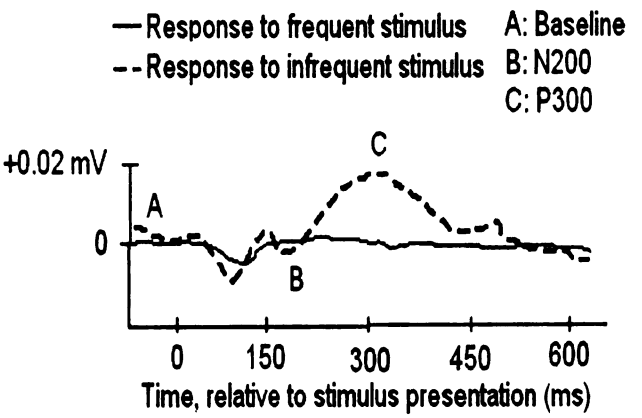


Figure 1.3. The P300 and N200 components can be seen in the averaged response to the rare stimuli from this experiment.

presented one by one to form a sentence, and the last word does not make sense. The worse fit a word is for the sentence, the stronger the response. It is only related to semantic, not grammatical errors. The response is weaker when the word is close semantically to what makes sense, i.e. "drink" instead of "eat." It has also been shown to be caused by metaphors, even when the subject grasps the metaphor.

The error-related negativity (ERN) is a negative component that appears when a subject makes an error in a response task to a stimulus. For example, the subject may have to respond with right or left hand to two types of visual or audial stimuli, this component will show up when the subject responds with right when he/she should respond with left, or vice versa. The ERN typically peaks around 150 ms after subject response. Amplitude of the ERN is larger with more emphasis on accuracy over speed. It also increases with an incorrect response differing from correct response by more movement parameters. The ERN can be seen in tasks where there is no error correction, as in a go-no go task, meaning it is involved in error detection as well as

probably error correction. It is not certain to what degree the ERN is involved in both detection and correction, though.

These ERP components are the peaks and troughs in the waveforms that covary in response to experiment manipulations. Each component is viewed to index some aspect of cognitive processing. The traditional view of components relating to the peaks and troughs is one way to define components. Another way would be as the aspects of the ERP that covary across subjects, conditions, or locations. A third definition of components are terms that directly correspond to the neural generating structures. The traditional representation of components is somewhat arbitrary, as these components may be represented as summations of the components found in these other ways. Finding the actual underlying sources is how to reduce the components to their most basic level. It is this way that the most insight could be gained about the processes of the brain.

The relationship between cognitive processes and ERP activity was studied with no reference to the underlying sources until the 90's. The problem of solving what will be measured at the electrodes given the sources in the brain is the forward problem, and so solving for the sources given only the readings of the electrodes is the inverse problem that must be solved.

1.6 Organization of Thesis

In Chapter 2, the background on common data reduction and blind source separation (BSS) techniques are presented and discussed. Chapter 3 introduces the problem of underdetermined blind source separation (UBSS), and a UBSS technique on the time-frequency plane is presented. In Chapter 4, the proposed UBSS method is applied to a multiple trial ERP data set to extract neuronal sources. The extracted neuronal sources are compared to ICA using measures of localization, sparsity, disjointness and variance. Chapter 5 concludes the thesis with a summary of contributions and

discussion of future work.

CHAPTER 2

SIGNAL PROCESSING METHODS FOR ERP COMPONENT EXTRACTION

2.1 Problem Statement

EEG/ERP signals are often assumed to be produced by distinct neuronal sources from distinct locations within the brain. These sources are conducted through the brain, skull, and scalp and create a potential measured by the electrodes. The readings at the electrodes are assumed to be instantaneous linear mixtures of the sources in noise following the model,

$$\mathbf{X} = \mathbf{AS} + \mathbf{V}, \tag{2.1}$$

with $\mathbf{X}$ representing the observation matrix, $\mathbf{A}$ the mixing matrix, $\mathbf{S}$ the source matrix, and $\mathbf{V}$ the noise matrix. The observation matrix, $\mathbf{X}$, is an $M \times p$ matrix with each row representing the reading at one electrode through time p, and each column, $\mathbf{x}(t)$ is the array of M sensor readings at time t. The source matrix, $\mathbf{S}$, is of size $N \times p$, where each row is the underlying source through time p, and each column, $\mathbf{s}(t)$ is the source array at time t. It is assumed that the noise matrix is negligible, the the representation becomes:

$$\mathbf{X} = \mathbf{AS} \tag{2.2}$$

If a linear transform is applied to the data, $\mathbf{X}$, the nature of the model is still the same,

$$\tilde{\mathbf{X}} = \mathbf{A}\tilde{\mathbf{S}}, \tag{2.3}$$

where $\tilde{\mathbf{X}}$ and $\tilde{\mathbf{S}}$ are the linearly transformed representations of $\mathbf{X}$ and $\mathbf{S}$, respectively. In the case that the quadratic time-frequency transform that will be used in this

thesis is applied to the data, the mixing model will be changed to,

$$\hat{\mathbf{X}} \approx \mathbf{A}^2\hat{\mathbf{S}} = \mathbf{B}\hat{\mathbf{S}}. \tag{2.4}$$

Here, $\hat{\mathbf{X}}$ and $\hat{\mathbf{S}}$ represent the quadratically transformed data and source matrices, respectively. For simplicity, equation 2.2 will always be considered the mixing equation, and it will be understood when using the quadratic transform, $\mathbf{X}$, $\mathbf{A}$, and $\mathbf{S}$ will actually represent $\hat{\mathbf{X}}$, $\mathbf{B}$, and $\hat{\mathbf{S}}$, respectively. A diagram of the mixing model is shown in Figure 2.1.

The model could not be written in this matrix form if the mixing process were assumed to be non-linear. Also, if the mixing matrix, $\mathbf{A}$, is assumed to be time-varying, a different matrix, $\mathbf{A}(t)$, would have to be specified for each $\mathbf{x}(t)$ and $\mathbf{s}(t)$. It is assumed that the mixing process is linear and time-invariant, therefore $\mathbf{A}$ is assumed to be constant, and is of size $M \times N$. Each mixing matrix entry, $[a_{ij}]$, represents the relative strength of source j at sensor i. This also assumes negligible delay between source activation and sensor reading. Otherwise a convolutive mixture would be necessary, in which the readings, $\mathbf{x}(t)$, would be a sum of different mixing matrices, $\mathbf{A}_0, \mathbf{A}_1, \ldots$, multiplied by $\mathbf{s}(t), \mathbf{s}(t-1) \ldots$, etc. The matrix $\mathbf{V}$ is additive noise with same size as $\mathbf{X}$.

With only the data collected from the electrodes available, $\mathbf{X}$, the goal is to extract the underlying sources by solving for $\mathbf{S}$. In general, a method must assume a certain structure for the underlying source signals and define a corresponding cost function for extracting the sources. In this chapter, some existing methods used in addressing this problem, i.e. data reduction and source separation, are presented.

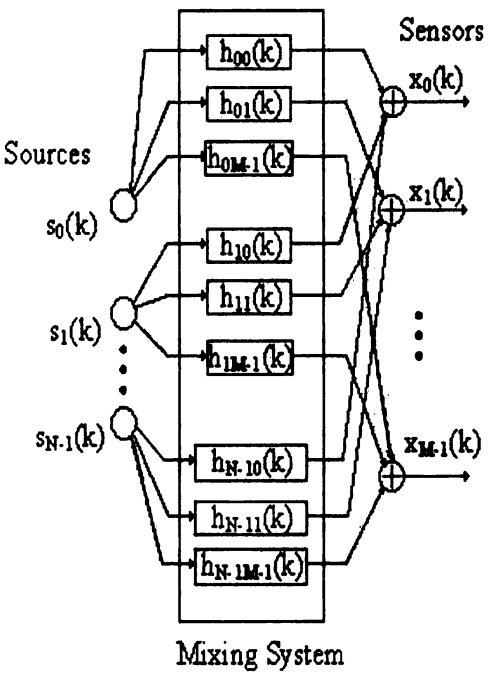


Figure 2.1. (a) The diagram of the mixing model with N sources and M electrodes (sensors). The transfer functions are simply scale factors since this is an instantaneous mixture.

2.2 Time-Frequency Distributions

Some of the methods introduced in this chapter depend on representing a signal in the joint time-frequency plane, so time-frequency distributions (TFDs) are introduced here. A TFD, $C(t,\omega)$, from Cohen's class can be expressed as *[16]:

$$C(t,\omega) = \int\int\int \phi(\theta,\tau)s(u+\frac{\tau}{2})s^*(u-\frac{\tau}{2})e^{j(\theta u-\theta t-\tau\omega)}du\,d\theta\,d\tau, \qquad (2.5)$$

*All integrals are from $-\infty$ to ∞ unless otherwise stated.

where $\phi(\theta, \tau)$ is the kernel function, and $s(t)$ is the signal.

The Short-Time Fourier Transform (STFT) is a simple instance of a TFD from Cohen's class, and can be represented as:

$$S_h(t, \omega) = \int s(\tau) h(\tau - t) e^{-j\tau\omega} d\tau. \tag{2.6}$$

The STFT involves performing Fourier Transforms on windowed data, where $h(t)$ is the window function and t represents its translation. A shorter duration of the window function, $h(t)$, provides higher time resolution, but at the expense of frequency resolution. Likewise, time resolution will suffer for increased frequency resolution. Thus, the representation of the data with the STFT is generally smeared. Also, the STFT representation does not necessarily maintain the time and frequency marginals, meaning the signal energy has been misrepresented [17].

The quadratic Wigner distribution is also of Cohen's class and is defined as follows:

$$W(t, \omega) = \int s(t + \frac{\tau}{2}) s^*(t - \frac{\tau}{2}) e^{-j\tau\omega} d\tau. \tag{2.7}$$

For quadratic TFDs, the cross-terms or interference occurs when the signal is multicomponent. The cross-terms correspond to the interaction between the different sources and do not contribute directly to the energy distribution of the individual sources, and thus are undesirable. For this reason, reduced interference distributions (RIDs) are used, designed using $\mid \phi(\theta, \tau) \mid \ll 1$ for $\mid \theta\tau \mid \gg 0$, that satisfy the energy preservation and the marginals [18].

Since the distributions will be implemented in discrete-time, the TFD can be

expressed as:

$$TFD(n,\omega;\psi) = \sum_{n_1=-N}^{N} \sum_{n_2=-N}^{N} x(n+n_1)x^*(n+n_2)\psi\left(-\frac{n_1+n_2}{2}, n_1-n_2\right) e^{-j\omega(n_1-n_2)}, \tag{2.8}$$

where ψ is the discrete-time kernel in the time and time-lag domain.

Cohen's class of distributions offer several advantages to other time-frequency analysis methods such as uniform time and frequency resolution, energy preservation and the marginals [16].

2.3 Data Reduction

Experimental ERP data is recorded over p time points per trial, with data available from multiple electrodes per trial, performed over multiple trials for multiple subjects. Each electrode reading per trial is represented as a p-dimensional observation if p is the number of time points measured and analyzed in time, or a P-dimensional observation if P is the number of total time-frequency points when analyzed in the time-frequency domain. If **J** is the number of subjects, **M** the number of electrodes, and **T** the number of trials, then the number of p- or P-dimensional observations to be analyzed is **J**×**M**×**T** and quickly becomes a large amount of data to process. Methods to reduce the data to a smaller number of meaningful components becomes an issue to effectively analyze the data.

2.3.1 Principal Component Analysis

Principal Component Analysis (PCA) is widely used in signal processing and pattern recognition applications. It is importantly used for data reduction with respect to ERP analysis. Like trial averaging, it is run over multiple trials, but PCA will itself extract multiple components, where averaging results in one waveform in which peaks are analyzed from. PCA reduces the ERP measurements into multiple orthogonal

components which covary over the trials as a result of experimental manipulations. The principal components (PCs) may be such that a summation over different PCs may correspond to a commonly known component, such as P300, but the PCs most likely do not represent the underlying sources.

With observations $\mathbf{x}(t) = [x_1(t), x_2(t), \ldots, x_M(t)]^T$, PCA is performed such that

$$y_i(t) = \sum_{k=1}^{M} w_{ki} x_k(t) = \mathbf{w}_i^T \mathbf{x}(t) \tag{2.9}$$

where y_i is the i^{th} principal component and $\mathbf{w}_i$ is a weight vector. The first principal component, y_1, is chosen so that its variance is maximized by choice of weight vector $\mathbf{w}_1$. The variance of y_1 is dependent on the orientation and norm of $\mathbf{w}_1$ and increases with the norm, so $\mathbf{w}_1$ is constrained to Euclidean l_2 norm of the weight vector, $||\mathbf{w}_1||_2 = 1$. The following y_i's are found to be orthogonal to all previous y_j, $j < i$, such that the variance is maximized in the subspace orthogonal to the space spanned by y_j, $j = 1, \ldots, i-1$. The solution to this problem is equivalent to eigenvalue decomposition of the data covariance matrix $\mathbf{R}$.

The data covariance matrix is defined as:

$$\mathbf{R} = \frac{1}{k} \sum_{t=1}^{k} \mathbf{x}(t)\mathbf{x}(t)^T. \tag{2.10}$$

The eigenvectors and eigenvalues of $\mathbf{R}$ are then found according to:

$$\mathbf{R} = \mathbf{V}\mathbf{\Lambda}\mathbf{V}^T \tag{2.11}$$

where $\mathbf{\Lambda}$ is the $M \times M$ diagonal matrix with entries representing the eigenvalues in decreasing order of magnitude $\lambda_1 > \lambda_2 > \cdots > \lambda_M$, and $\mathbf{V} = [\mathbf{v}_1, \mathbf{v}_2, \ldots, \mathbf{v}_M]$ with columns representing the corresponding eigenvectors. The first m eigenvectors are

kept and used such that $\mathbf{w}_1 = \mathbf{v}_1, \mathbf{w}_2 = \mathbf{v}_2, \ldots, \mathbf{w}_m = \mathbf{v}_m$, and the variances of the principal components y_i are given by the respective eigenvalues λ_i. The resulting approximation to the original data set is:

$$\hat{\mathbf{x}}(t) = \sum_{i=1}^{m} \lambda_i \mathbf{v}_i(t). \tag{2.12}$$

A threshold is chosen such that the number of principal components kept, m, is determined. The error between $\mathbf{x}(t)$ and $\hat{\mathbf{x}}(t)$ goes to zero as m approaches M, and the eigenvalues, $\sum_{i=m+1}^{M} \lambda_i$ represent the error. A common decision rule is to set a minimum value for each eigenvalue such that eigenvector $\mathbf{v}_i$ is kept if $\lambda_i > T$, where T is the threshold. Another common decision rule is to keep the eigenvectors that together explain a minimum amount of variance. That is, such that $\frac{\sum_{i=1}^{m} \lambda_i}{\sum_{k=1}^{M} \lambda_k} > T$, where T is the threshold.

To be viewed as source extraction, the sources would have to be assumed to be orthogonal in time. This is not an assumption that is often made, so this method is not used to determine the underlying sources. Its value comes in data reduction, so that further analysis techniques can more efficiently be applied to the data by using the extracted components.

2.3.2 Matching Pursuit

Matching Pursuit (MP) is an algorithm that decomposes a signal linearly into a set of functions, called time-frequency atoms, from an overcomplete dictionary. It can be used in simplifying the representation of a particular signal via approximation.

The dictionary is composed of functions, $g(t)$, scaled, translated, and modulated as follows:

$$g_\gamma(t) = \frac{1}{\sqrt{s}} g(\frac{t-u}{s}) e^{j\xi t}, \tag{2.13}$$

where s is scale, u translation, ξ modulation, and $\gamma = (s, u, \xi) \in \Gamma$. The $\frac{1}{\sqrt{s}}$ term

normalizes $g_\gamma(t)$. The dictionary, $\mathbf{D} = [g_\gamma]_{\gamma \in \Gamma}$, is overcomplete in that it provides more than a sufficient basis for the signals it is to represent.

The first step in MP is choosing a function, g_{γ_0}, from the given dictionary, $\mathbf{D}$. This g_{γ_0} is projected on the original signal, f, and $R^1 f$ is defined as the residue such that:

$$f = \langle f, g_{\gamma_0} \rangle g_{\gamma_0} + R^1 f \tag{2.14}$$

and because the residue is orthogonal to the dictionary function,

$$||f||^2 = |\langle f, g_{\gamma_0} \rangle|^2 + ||R^1 f||^2. \tag{2.15}$$

The function, g_{γ_0}, must be chosen to maximize $|\langle f, g_{\gamma_0} \rangle|$ so that the residual, $R^1 f$, is minimized.

The second step is then to choose g_{γ_1} from $\mathbf{D}$. This is found by maximizing $|\langle R^1 f, g_{\gamma 0} \rangle|$ so that the new residue is $R^2 f$ and $R^1 f = \langle R^1 f, g_{\gamma 0} \rangle g_{\gamma 0} + R^2 f$.

These iterations are repeated likewise so at the m^{th} iteration, the signal approximation is:

$$f = \sum_{n=0}^{m-1} \langle R^n f, g_{\gamma n} \rangle g_{\gamma n} + R^m f, \tag{2.16}$$

where $R^0 = f$. As the iteration number, $m \to \infty$, $||R^m f||^2 \to 0$. The algorithm should be run until $||R^m f||^2 < T$, for a predetermined threshold, T.

The MP algorithm can offer compact representations of signals under the right conditions, however this is dependent on the dictionary chosen. A larger library of atoms can be chosen to better represent the dictionary at the expense of computational complexity, but signal components must be reasonably approximated by dictionary atoms. Frequency-modulated Gaussian functions, called gabor logons, give optimal joint time-frequency localization [20], and therefore are well suited in representation of ERPs. MP is a greedy algorithm, so it does not necessarily provide an

optimal solution.

2.3.3 Principal Component Analysis and Matching Pursuit on Time-Frequency

Using a frequency transform like the Fourier Transform obscures the time localization of the signal, while analysis in time domain loses information about the frequencies that constitute the signal. It has been shown that using a TFD for EEG data can produce results not available with conventional analysis techniques [21]. Using a TFD of Cohen's class has the advantage over a wavelet transform (WT) that it has constant time-frequency atoms whereas the WT suffers from low time resolution at low frequencies and low frequency resolution at higher frequencies. Under noisy conditions, the RID was shown to provide better separation of signal components than using the WT [17]. The RID is chosen for signal representation for these advantages. Because of the two-dimensional nature of a signal under a TFD, there is much more data to be processed than when analyzing signals in time or frequency alone. It was shown that data reduction using PCA on TFD representations of EEG data can provide meaningful components [17].

This approach assumes time-frequency stationarity of sources. If a single source shifts in the time-frequency plane over multiple trials, either multiple components explaining the same source will be extracted, or the component representing the source will be spread to cover the area of source shift.

Like in the case of time domain, to be viewed as source extraction, the sources would have to be assumed to be orthogonal in the time-frequency plane. So again, its value comes in data reduction, so that further analysis techniques can more efficiently be applied to the data by using the extracted components.

Given that PCA can successfully be used in data reduction and that the extracted components are localized in the time-frequency plane, it has been shown possible to meaningfully reduce the amount of data further [26]. Using a dictionary of Gabor

logons, a Matching-Pursuit (MP) algorithm is applied to components extracted with PCA on TFDs. The resulting components were shown to represent the meaningful variance of the original data well. Because the components are reduced to a few time and frequency parameters, the data is reduced further than using the time-frequency surfaces of the principal components.

2.4 Blind Source Separation

2.4.1 Independent Component Analysis

In recent years, multivariate data analysis involving decomposing the measurements into several independent time series has become a popular way to describe the 'source' signals in the brain. Independent Component Analysis (ICA) solves for the unmixing matrix $\mathbf{W} = \mathbf{A}^{-1}$ such that the number of electrodes is at least as large as the number of underlying sources ($M \geq N$), $\mathbf{W}$ is full rank, the estimated sources, $\mathbf{Y} = \mathbf{WX}$, are as independent as possible [22], and $\mathbf{A}$ is as described in section 2.1.

The Infomax algorithm introduced by Bell and Sejnowski [23] runs to maximize the mutual information that the system output contains about its input. In its application to ERPs, the estimated brain sources represent the output and the electrode readings represent the input. Consider $\mathbf{x}$ to be the input vector and $\mathbf{y}$ the output vector where $\mathbf{y} = g(\mathbf{Wx} + \mathbf{w}_0)$ with $\mathbf{W}$ a mixing matrix and $\mathbf{w}_0$ a bias vector. The multivariate pdf of $\mathbf{y}$ is:

$$f_{\mathbf{y}}(\mathbf{y}) = \frac{f_{\mathbf{x}}(\mathbf{x})}{|J|} \tag{2.17}$$

and

$$J = det \begin{bmatrix} \frac{\partial y_1}{\partial x_1} & \cdots & \frac{\partial y_1}{\partial x_n} \\ \vdots & & \vdots \\ \frac{\partial y_n}{\partial x_1} & \cdots & \frac{\partial y_n}{\partial x_n} \end{bmatrix}. \tag{2.18}$$

The output entropy is $H(\mathbf{y}) = -E[ln f_{\mathbf{y}}(\mathbf{y})]$. By plugging equation 2.17 into this

equation the entropy becomes $H(\mathbf{y}) = E\left[ln|J|\right] - E[lnf_{\mathbf{x}}(\mathbf{x})]$. The term on the right, the entropy of $\mathbf{x}$, is unaffected by $\mathbf{W}$, so that maximization of $ln|J|$ is required to maximize $|H(\mathbf{y})|$. Gradient ascent learning rules are derived as:

$$\Delta\mathbf{W} \propto [\mathbf{W}^T]^{-1} + (\mathbf{1} - 2\mathbf{y})\mathbf{x}^T \tag{2.19}$$

$$\Delta\mathbf{w}_0 \propto \mathbf{1} - 2\mathbf{y}, \tag{2.20}$$

and the rule for individual weights is therefore:

$$\Delta w_{ij} \propto \frac{cofw_{ij}}{det\mathbf{W}} + x_j(1 - 2y_i), \tag{2.21}$$

where cof is the cofactor and det is the determinant. This gradient ascent algorithm is efficient if the pdf's of the inputs are super-gaussian with no more than one source being gaussian, and when the mixing matrix is not almost singular.

The process of Principal Component Analysis (PCA) can be compared with that of ICA. They are similar, however PCA seeks out sources which are orthogonal to each other, only relying on second order statistics, while ICA uses the less restrictive assumption of independence and uses higher order statistics [24]. In PCA, each progressive component explains as much variance of the data as possible, while the variance of each component in ICA is more spread out [27].

In using ICA, it is required to have sufficient data points to represent independence. This calls for a few times more data points than there are electrodes. It is required to have as many electrodes as there are underlying sources in order to successfully extract the independent sources. Also, the sources will only be successfully separated given that the assumption of independence is true. This assumption of independence between multiple sources in the brain is not necessarily true.

2.4.2 Blind Source Separation on the Time-Frequency Plane

Much work is being done in BSS based on the assumptions of disjointness and sparsity of the sources. These techniques do not assume independence of the sources, and they are designed for the underdetermined case, that is, when there are more sources than sensors, as described in section 3.1. Time or frequency only representations do not provide sufficient sparsity, so use of a time-frequency transform is necessary to assume sparse sources.

A method presented in [28] uses the time-frequency representation of the short-time Fourier transform (STFT) for BSS. This method was proposed for speech signals and uses masking to separate the sources. The sources are assumed to be approximately disjoint, that is approximately non-overlapping, in the time-frequency domain. Masks are created using magnitude and phase information of the mixtures. Since the assumption that time lag is negligible for EEG data is made, the phase information can be ignored in application to EEG data, and in this way the method can be extended to other time-frequency representations. This algorithm can extract more sources than sensors, but only considers the case in which two mixtures are provided.

Sources $s_1(t)$ through $s_n(t)$ are represented in both mixtures $x_1(t)$ and $x_2(t)$ at different scales because of their different locations. The scales in the first mixture are considered to be one, such that the mixtures can be represented as follows:

$$x_1(t) = s_1(t) + s_2(t) + \cdots + s_n(t), \tag{2.22}$$

$$x_2(t) = a_1 s_1(t) + a_2 s_2(t) + \cdots + a_n s_n(t). \tag{2.23}$$

The STFT of the mixtures are represented as follows:

$$X_1(t,\omega) = S_1(t,\omega) + S_2(t,\omega) + \cdots + S_n(t,\omega), \tag{2.24}$$

$$X_2(t,\omega) = a_1 S_1(t,\omega) + a_2 S_2(t,\omega) + \cdots + a_n S_n(t,\omega). \tag{2.25}$$

The ratio R_{21} is defined as

$$R_{21} = \frac{X_2(t,\omega)}{X_1(t,\omega)}. \tag{2.26}$$

In this way, if the sources are completely time and frequency disjoint, only one source will be active at any point in R_{21} and its value will be that of the scale of the active source. All points at which a given source is active will result in the same ratio of R_{21} giving that source's scale. If source m is active, then $R_{21} = a_m$, and by creating a mask for each different value in R_{21}, the sources can be separated from the original mixtures. If the sources are only approximately time and frequency disjoint, then one source is dominant at any point. If the values of R_{21} are plotted in a histogram, then there will be a peak for each source, the peaks will be located at approximately the scale factor of each source, and all values near should belong to that source.

The second approach to BSS uses the nonlinear time-frequency distributions. Blind source separation based on spatial time-frequency distributions achieve separation by joint diagonalization of the auto-terms in the spatial time-frequency distributions [29, 30, 31, 32].

The discrete-time implementation of Cohen's class TFD for signal $x_1(t)$ is[16]:

$$D_{x_1 x_1}(t,\omega) = \sum_{l=-\infty}^{\infty} \sum_{m=-\infty}^{\infty} \psi(m,l) x_1(t+m+l) x_1^*(t+m-l) e^{-j2\omega l}, \tag{2.27}$$

with t being time index and ω frequency index. The cross-TFD of two signals $x_1(t)$ and $x_2(t)$ is then defined by:

$$D_{x_1 x_2}(t,\omega) = \sum_{l=-\infty}^{\infty} \sum_{m=-\infty}^{\infty} \psi(m,l) x_1(t+m+l) x_2^*(t+m-l) e^{-j2\omega l}. \tag{2.28}$$

Equations 2.27 and 2.28 are used to define the spatial time-frequency distribution

(STFD) matrix as:

$$D_{\mathbf{xx}}(t,\omega) = \sum_{l=-\infty}^{\infty} \sum_{m=-\infty}^{\infty} \psi(m,l)\mathbf{x}(t+m+l)\mathbf{x}^*(t+m-l)e^{-2\omega l}. \tag{2.29}$$

In [29], the method is a two-step process: first the mixing matrix $\mathbf{A}$ is transformed into a unitary matrix $\mathbf{U}$ via whitening, second $\mathbf{U}$ is retrieved through joint diagonalization of a set of whitened data STFD matrices. Given that all source signals $s_i(t)$ are mutually uncorrelated, $\mathbf{W}$ can be determined from $\mathbf{R}$ as defined in 2.10 through:

$$\mathbf{W}(\mathbf{R} - \sigma^2\mathbf{I})\mathbf{W}^T = \mathbf{W}\mathbf{A}\mathbf{A}^T\mathbf{W}^T = \mathbf{I}, \tag{2.30}$$

where σ^2 is noise variance. The whitened STFD matrix is then solved with:

$$\mathbf{D}_{\mathbf{zz}}(t,\omega) = \mathbf{W}\mathbf{D}_{\mathbf{xx}}(t,\omega)\mathbf{W}^T = \mathbf{U}\mathbf{D}_{\mathbf{SS}}(t,\omega)\mathbf{U}^T, \tag{2.31}$$

with $\mathbf{z}(t) = \mathbf{W}\mathbf{x}(t)$ the whitened data vector and $\mathbf{U} = \mathbf{W}\mathbf{A}$ is unitary. The source signals are estimated such that $\mathbf{s}(t) = \mathbf{U}^T\mathbf{W}\mathbf{x}(t)$.

This approach facilitates separation of Gaussian sources with identical spectral shapes but different time-frequency localization properties. However, this method require extensive computing and a priori knowledge about the structure of the signals.

Bofill and Zibulevsky developed a two stage approach to solving this problem in [33] under the model of equation 2.1, specifically for the case of two mixtures, $M = 2$. The data is represented in the STFT domain to improve sparsity over representation in time. Let $x_1(t)$ and $x_2(t)$ represent the measurements at each respective sensor at time t. When sources are sparse, most sources contribute almost zero to the measured mixtures at any measured point. When one source, $s_j(t)$, is strong, then, it is likely the remaining sources are close to zero, and the measured vector, $\mathbf{x}(t)$, lines up in the direction of the column of the mixing matrix, $\mathbf{a}_j$. In the two mixture case, a scatter

plot of $x_1(t)$ vs. $x_2(t)$ will show a distinct direction for each source if the sources are sufficiently sparse, and are mixed in independent directions. Using polar coordinates, each measured data point is given a radius, l_t, and angle, θ_t. They define a global potential function:

$$\Phi(\theta, \lambda) = \sum_t \phi(\lambda(\theta - \theta_t)), \tag{2.32}$$

with

$$\phi(\alpha) = \begin{cases} 1 - \frac{\alpha}{\pi/4}, \text{ for } |\alpha| < \pi/4 \\ 0, \text{ elsewhere} \end{cases} . \tag{2.33}$$

Evaluating this equation around all values of θ will give peaks along the angles of $\mathbf{a}_j$. The radius, l_t, acts as a weight to count more reliable data for more, and λ adjusts angular width of the peaks. The number of peaks found serves as the estimate of number of sources and columns of the mixing matrix. The mixing matrix is then used to estimate the underlying sources following the same procedure outlined in section 3.4.

Proposed in [34] is a two stage approach to BSS. It is an iterative algorithm that first estimates the mixing matrix, and from this estimates the original sources. It is two stages because it assumes an underdetermined case, that is there are more sources than there are sensors. The mixtures are examined in a wavelet packet domain in hopes to ensure sparsity.

Similar to the masking technique, this algorithm relies on sources being approximately disjoint. First, the wavelet packet coefficients of observation matrix $\mathbf{X}$ are calculated as $\tilde{\mathbf{X}}$, so that there are n observations and each row represents the wavelet packet representation of one observation and is of dimension N. Second, a ratio

matrix is created using the wavelet packets coefficients,

$$\tilde{\tilde{\mathbf{X}}} = \begin{bmatrix} \frac{\tilde{x}_1(1)}{\tilde{x}_q(1)} & \cdots & \frac{\tilde{x}_1(N)}{\tilde{x}_q(N)} \\ \vdots & \vdots & \vdots \\ \frac{\tilde{x}_n(1)}{\tilde{x}_q(1)} & \cdots & \frac{\tilde{x}_n(N)}{\tilde{x}_q(N)} \end{bmatrix}, \tag{2.34}$$

with $q \in 1, \ldots, n$. A submatrix of $\tilde{\tilde{\mathbf{X}}}$ is then found:

$$\begin{bmatrix} \frac{\tilde{x}_1(i_1)}{\tilde{x}_q(i_1)} & \cdots & \frac{\tilde{x}_1(i_K)}{\tilde{x}_q(i_K)} \\ \vdots & \vdots & \vdots \\ \frac{\tilde{x}_n(i_1)}{\tilde{x}_q(i_1)} & \cdots & \frac{\tilde{x}_n(i_K)}{\tilde{x}_q(i_K)} \end{bmatrix} = \begin{bmatrix} \frac{a_{11}}{a_{q1}} & \cdots & \frac{a_1 1}{a_{q1}} \\ \vdots & \vdots & \vdots \\ \frac{a_{n1}}{a_{q1}} & \cdots & \frac{a_n 1}{a_{q1}} \end{bmatrix}. \tag{2.35}$$

Ideally, this submatrix will have identical columns, which would represent one column of the mixing matrix. This would be without the presence of noise. In real conditions, noise is present, and so the submatrix has approximately identical columns, and the mean of these columns represents the estimation of a column of the mixing matrix. Another such submatrix of $\tilde{\tilde{\mathbf{X}}}$ is found to have a different set of approximately identical columns, and is used to estimate another column of the mixing matrix. This process is repeated, each time estimating a column of the mixing matrix, until there are no more unique submatrices of $\tilde{\tilde{\mathbf{X}}}$. Next, q is incremented to find a new $\tilde{\tilde{\mathbf{X}}}$ and the process repeats until $q = n$. All estimated columns of the mixing matrix are put together for its estimation. Redundant columns of the mixing matrix are eliminated and a final estimate of the mixing matrix is achieved. After this linear programming can be used to estimate the sources, an approach to this is outlined in section 3.4.

Another two stage approach like this is proposed in [35]. This approach assumes the underdetermined case as well as sparsity. It is also proposed to use wavelet packet in this algorithm. To estimate the mixing matrix, K-means clustering is used. The implementation of the K-means clustering is much simpler than the iterations used

in [34]. This algorithm is described in Chapter 3.

The method proposed in this thesis is an extension of the algorithm from [35]. Instead of working in the wavelet packet domain, it is proposed to work in the higher resolution TFD of Cohen's class. It is assumed that the added sparsity of using the TFD instead of wavelet packet will provide better results and make unnecessary a more complex algorithm.

CHAPTER 3

UNDERDETERMINED SOURCE SEPARATION IN TIME-FREQUENCY DOMAIN

An extension of the underdetermined blind source separation method proposed in [35] is presented here. An overview of underdetermined blind source separation is given first. Next, the two-stage approach used to solve the BSS problem in the underdetermined case is presented. The first stage is to estimate the mixing matrix, and the second is to estimate the sources.

3.1 Underdetermined Blind Source Separation

Underdetermined Blind Source Separation (UBSS) is a special case of the mixing model presented in section 2.1 in which the number of sources is more than the number of sensors, $\mathbf{N} > \mathbf{M}$. In this case the estimation of the mixing matrix alone is not enough for the estimation of the underlying sources since the mixing matrix will not be invertible like in the determined case. It is necessary to make assumptions about the data to overcome this, and increasingly popular is assuming that the sources are sparse in a particular representation [19, 36, 37]. For a source to be sparse, most data points must be close to zero, so the pdf of its values must peak at zero and have long tails. Because of this, the Laplacian distribution is often used to model the pdf of a sparse source [33]. Representing the sensor readings in the time or frequency domain alone does not often induce sparsity, so use of wavelet packets and TFDs are often employed. It was shown in [38] that use of TFDs can provide more robust separation under noisy conditions, so TFDs are used in this application. If the sources are sparse, then they are also more likely to be non-overlapping, or approximately disjoint, which can be used to make estimation of the mixing process easier.

3.2 Determination of the Mixing Matrix

The mixing matrix to be solved for is $\mathbf{A}$ in

$$\mathbf{X} = \mathbf{AS}, \tag{3.1}$$

where $\mathbf{X}$ is the $M \times P$ observation matrix with each row being the TFD of one electrode, $\mathbf{S}$ is the $N \times P$ source matrix with each row being the TFD of one source, and each column of both $\mathbf{X}$ and $\mathbf{S}$ represents all values at one joint time-frequency point. The mixing matrix actually represents the element-by-element square of the mixing matrix respresenting this mixing problem in the time domain. The number of sources is assumed greater than the number of electrodes, $N > M$.

The estimation of the mixing matrix relies on a representation of the data that renders the sources sufficiently sparse. In [35], it was proposed to work with the wavelet packets (WP) representation to provide sparsity of sources. We propose to work with a TFD representation because of the increased time and frequency resolution of the data, ensuring a sparser representation. Use of WP and TFD were compared in [38], and these results are shown in section 3.5.

Due to the sparsity of the source signals in the time-frequency domain, it is likely that there exist many columns of $\mathbf{S}$ with only one nonzero entry. For instance, suppose that $\mathbf{s}_{1_j}, \cdots, \mathbf{s}_{K_j}$ are K columns of $\mathbf{S}$, where only the jth entry of each of these columns is nonzero. For this case we assume that j will mean the first entry of each column is nonzero. Then it follows

$$\mathbf{A}\mathbf{s}_{i_j} = \mathbf{a}_1 s_{1i_j} \quad i = 1, \cdots, K, \tag{3.2}$$

with s_{1i_j} representing the first element of column s_{i_j} since it is the only active element

in that column, and

$$[\mathbf{x}_{1_j}, \cdots, \mathbf{x}_{K_j}] = \mathbf{A}[\mathbf{s}_{1_j}, \cdots, \mathbf{s}_{K_j}] = [\mathbf{a}_1 s_{11_j}, \cdots, \mathbf{a}_1 s_{1K_j}], \tag{3.3}$$

where, $\mathbf{x}_{i_j}$ is the i_jth column of $\mathbf{X}$ corresponding to $\mathbf{s}_{i_j}$, $\mathbf{a}_1$ is the first column of $\mathbf{A}$, and $\mathbf{s}_{1i_j}$ is the first entry of $\mathbf{s}_{i_j}$. From equation (3.3), we see that each $\mathbf{x}_{i_j}$ is equal to $\mathbf{a}_1$ multiplied by a scalar s_{1i_j}, which means that these K column vectors of $\mathbf{X}$, $\mathbf{x}_{1_j}, \cdots, \mathbf{x}_{K_j}$, are distributed along the direction of $\mathbf{a}_1$. Thus, ideally after normalization, $\mathbf{x}_{1_j}, \cdots, \mathbf{x}_{K_j}$ are mapped to a unique vector on the multidimensional unit circle which is equal to $\mathbf{a}_1$. However, in practice, the sources are likely only approximately disjoint. That is, $s_{1_j}, \ldots, s_{K_j}$ are K columns of $\mathbf{S}$ with the j^{th} entry dominant, so that $s_{ji_j} >> s_{pi_j}$, $p \neq j$, where j is constant. When more than one source is non-zero, $\mathbf{x}_{1_j}, \cdots, \mathbf{x}_{K_j}$ are not exactly in the same direction as $\mathbf{a}_1$, but rather in the neighborhood of $\mathbf{a}_1$. This means that $\mathbf{a}_1$ lies at the center of $\mathbf{x}_{1_j}, \cdots, \mathbf{x}_{K_j}$.

Therefore, the K-means clustering method presented in the following section can be used to cluster the column vectors of the mixture matrix $\mathbf{X}$ into multiple clusters, where the center of each cluster corresponds to one column vector of the mixing matrix $\mathbf{A}$. By doing so, an estimate of the mixing matrix $\mathbf{A}$ is obtained, where each column, $\mathbf{a}_i$, is estimated by one of the resulting cluster centers.

3.3 K-means Clustering

K-means clustering is an iterative algorithm that seeks to minimize a squared-error criterion function in order to separate a completely unknown set of data into k different groupings [39]. Suppose that $\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_n$ are vector observations in a data set and make up realizations of k different distributions of random variables. Then $\mu_1, \mu_2, \ldots, \mu_k$ are the mean vectors of these distributions, and k-means seeks to cat-

egorize the observations, $\mathbf{x}_i$, into one of the k distributions such that the squared Euclidean distance, $||\mathbf{x}_i - \mu_j||^2$, is minimized. However, since the properties of the dataset are unknown, $\mu_1, \mu_2, \ldots, \mu_k$ must be estimated first, as $\hat{\mu}_1, \hat{\mu}_2, \ldots, \hat{\mu}_k$.

As a starting point, k random samples of the data are chosen as the initial mean estimates, $\hat{\mu}_j$. The distributions are then estimated by classifying all points, $\mathbf{x}_i$, into the group whose estimated mean it is closest to in the squared Euclidean sense, so that $\mathbf{x}_i \in \hat{\mu}_j$ when j is subject to

$$\min_j ||\mathbf{x}_i - \hat{\mu}_j||^2. \tag{3.4}$$

Once all data points are classified, the mean of each group is recalculated. Suppose m_j is the number of data points in the jth distribution, and $\mathbf{x}_{1_j}, \mathbf{x}_{2_j}, \ldots, \mathbf{x}_{m_j}$ are all data points. The new mean is then calculated as $\hat{\mu}_j = \frac{1}{m_j} \sum_{i=1}^{m_j} \mathbf{x}_{i_j}$. This process is repeated until convergence, when the estimated means do not change upon further iterations.

3.4 Estimation of the Source Signals for a Given Mixing Matrix

After obtaining the estimated mixing matrix, the next stage is to estimate the source signals. For a given mixing matrix $\mathbf{A}$ in equation (2.2), the source signals can be estimated by maximizing the posterior distribution $P(\mathbf{S}|\mathbf{X}, \mathbf{A})$ of $\mathbf{S}$. In general, there is no unique solution to this problem. Maximizing the posterior distribution is then done such that a maximally sparse set of sources are found. Under the assumption that the prior is Laplacian, maximizing posterior distribution can be implemented by solving the following optimization problem [40]:

$$\min \sum_{i=1}^{N} \sum_{j=1}^{P} |s_{ij}|, \quad \text{subject to } \mathbf{AS} = \mathbf{X}, \tag{3.5}$$

with N the number of sources, and P the number of time-frequency points. Hence, the l_1-norm

$$J_1(\mathbf{S}) = \sum_{i=1}^{N} \sum_{j=1}^{P} |s_{ij}| \tag{3.6}$$

can be used as the sparsity measure. The l_1 norm is preferred to l_0 norm, which is the actual level of sparsity, because the optimization is NP hard for l_0 norm but can be solved easily for l_1 norm using linear programming.

It is not difficult to prove that the optimization problem (3.5) is equivalent to the following set of P smaller scale linear programming (LP) problems:

$$\min \sum_{i=1}^{N} |s_{ij}|, \quad \text{subject to } \mathbf{A}\mathbf{s}_j = \mathbf{x}_j \quad \text{for } j = 1, \cdots, P. \tag{3.7}$$

In this way, the contribution to the column of $\mathbf{X}$ at one point of time-frequency should be dominated by one source in order to minimize the l_1-norm, giving the sparse solution.

Finally, we propose the following algorithm for estimating the source signals:

Algorithm:

1. Using the collected data $[z_1(t), z_2(t), \ldots, z_M(t)]^T$, obtain M TFDs and vectorize each to obtain the TFD mixture $\mathbf{X}$.

2. Normalize the column vectors of the TFD mixture $\mathbf{X}$ to obtain $\widehat{\mathbf{X}}$.

3. Take a sufficiently large positive integer k as the number of clusters, also the number of sources to estimate. Choose the initial points of iteration and the distance measure criterion. In this thesis, the squared Euclidean distance is chosen as the criterion.

4. Do K-means clustering on $\widehat{\mathbf{X}}$ followed by normalization to estimate the sub-optimal mixing matrix $\mathbf{A}$.

5. Using the estimated mixing matrix $\mathbf{A}$ and the mixtures $\mathbf{X}$, estimate the time-frequency representations $\mathbf{S}$ by solving the set of LP problems in equation (3.7).

The result is a matrix, $\mathbf{S}$, in which each row represents a vectorized TFD of an extracted sparse component. The number of sources to be extracted is defined by the user, and so a sufficiently large number must be chosen. It is shown in [35] that if k is chosen larger than the actual number of sources, the "extra" extracted sources appear as spurious noise sources, and can be ignored. The l_1 norm is used as a sparseness measure to ensure a unique solution, and the solution will be reliable if the sources are approximately disjoint.

3.5 Comparison Between Wavelet Packets and Time-Frequency Distributions

In this section, several examples will be used to illustrate the effectiveness of the proposed approach to separate the sparse source signals from their fewer mixtures in the time-frequency domain. The binomial kernel [16] is used for computing the TFD since it belongs to the class of reduced interference distributions (RIDs).

Example 1: The set of observed signals are two linear combinations of four Gabor logons. These four Gabor logons are centered at the time sample point and the normalized frequency (30,0.7), (160,-0.7), (70,-0.4), and (120,0.1), respectively. Each observed signal is first transformed to the time-frequency domain with $I = 50$ time samples and $L = 64$ frequency samples. Each TFD is then vectorized to form a TFD mixture matrix $\mathbf{X} = [\mathbf{X}_1; \mathbf{X}_2]$ of size 2×3200.

Figure 3.1 presents a scatter plot of the mixtures $\mathbf{X}$ ($\mathbf{X}_2$ vs. $\mathbf{X}_1$) in the time-frequency domain. It can be seen from this plot that almost all significant data points are distributed along four different directions, thus providing very good separability. The separation results using the proposed approach are illustrated in Figure 3.2. Figure 3.2(a) and (b) represent the two mixtures. The four extracted Gabor logon

signals are shown in Figure 3.2(c), (d), (e), and (f). The results indicate that these four Gabor logons can be successfully separated from their two mixtures using the proposed approach based on their sparsity with an average signal to interference ratio (SIR) of 36.1251 dB.

Example 2: Two mixtures of a chirp signal and two Gabor logons are given. The chirp signal has a linear frequency increasing from an initial normalized frequency of -0.2 to a normalized frequency of 0.2. The Gabor logons are the first two Gabor logons given in Example 1. A scatter plot of the two mixtures in Figure 3.3 shows that it is similar to the first example in that the distributions of data points belonging to different sources are along three different directions. Since the chirp signal overlaps with the two Gabor logons in the time domain, typical time domain separation methods can not be used to perfectly recover them. However, it is illustrated in Figure 3.4 that these three signals can be effectively extracted in the time-frequency domain using the proposed method with an average SIR of 32.7634 dB.

Example 3: In this example, the same two mixtures of four Gabor logons given in Example 1 are used. The effectiveness of the presented approach is compared for TFDs and wavelet packets (WP) in the presence of noise. Haar wavelet with five levels is used for the wavelet packet decomposition.

To show the effect of increased sparsity of TFDs, the mixtures at different levels of white Gaussian noise are considered. 100 Monte Carlo simulations are used for each noise level. The average mean squared error (MSE) between the extracted sources and the original sources is calculated for both the TFD and WP. The TFD and WP provide similar results when there is no noise. However, as the noise level increases, the performance of the WP rapidly degrades compared to that of the TFD. The MSE versus the signal-to-noise ratio (SNR) is shown in Figure 3.5 for both the TFD and WP. This result shows that the RID results in a more sparse time-frequency surface compared to the WP, which improves the robustness of BSS under noise.

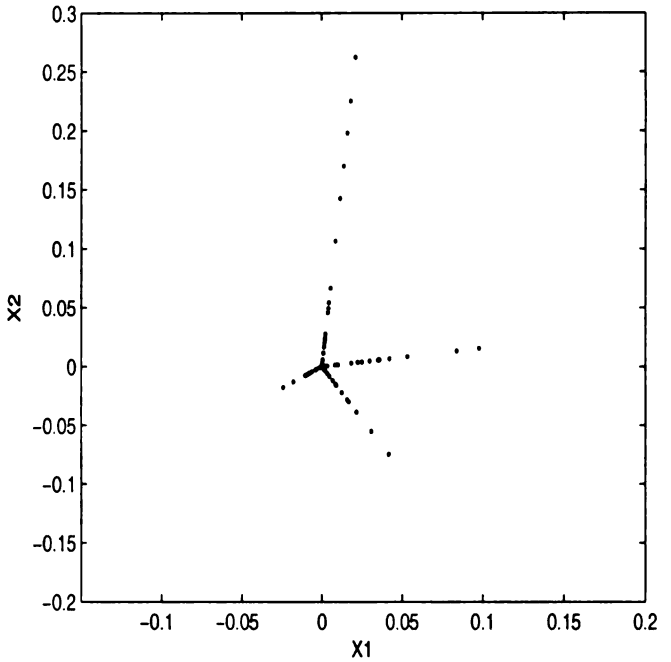


Figure 3.1. Scatter plot of two mixtures of four Gabor logons in the time-frequency domain

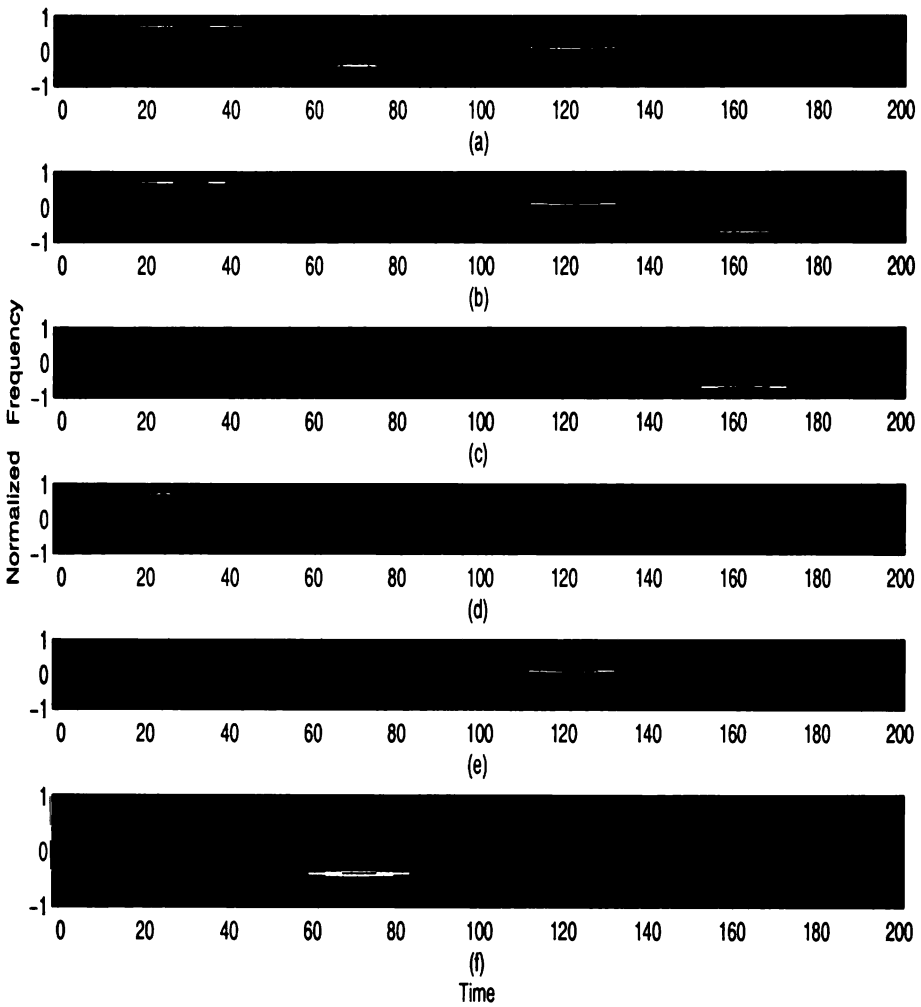


Figure 3.2. The mixtures and the separation of four Gabor logons: (a) and (b) two mixtures; (c), (d), (e), and (f) four extracted Gabor logons

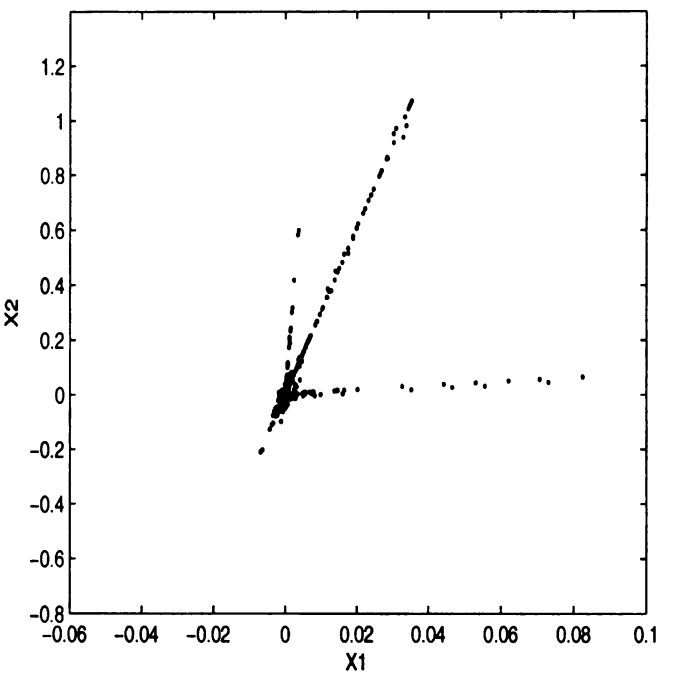


Figure 3.3. Scatter plot of two mixtures of a chirp and two Gabor logons in the time-frequency domain

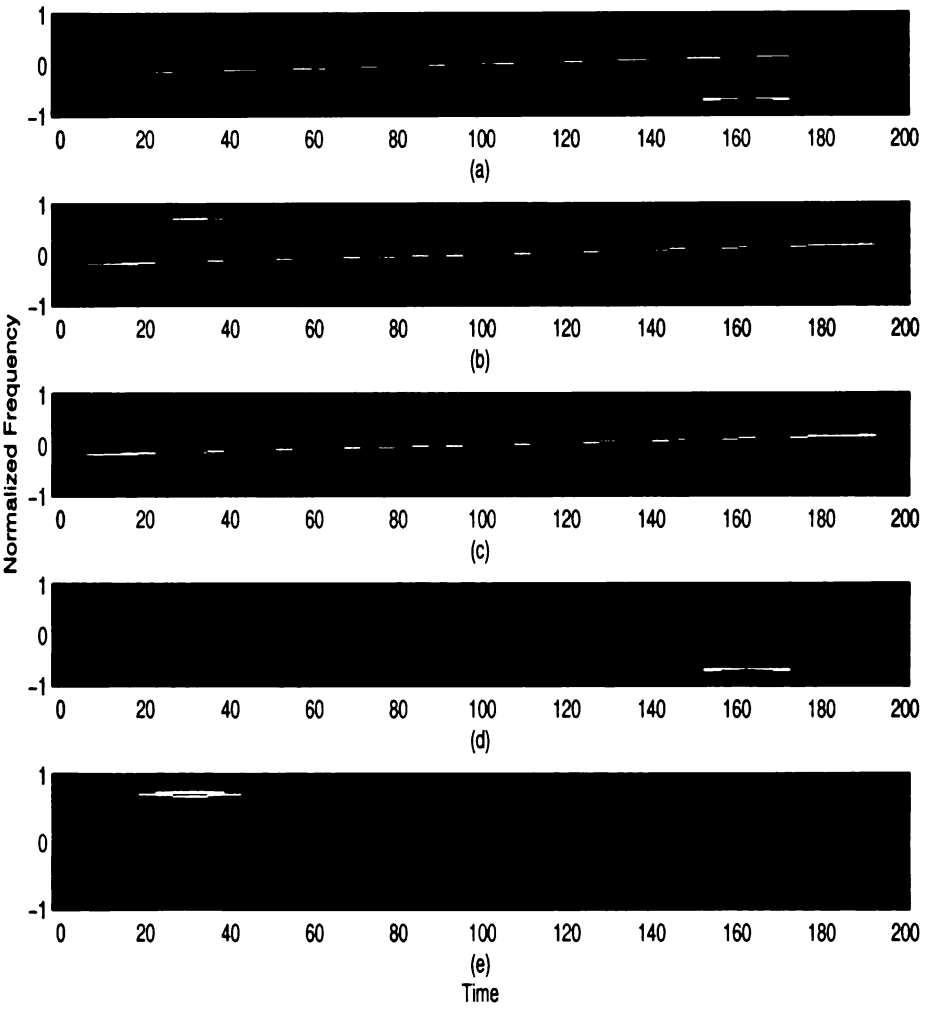


Figure 3.4. The mixtures and the separation of a chirp and two Gabor logons: (a) and (b) two mixtures; (c) extracted chirp; (d) and (e) two extracted Gabor logons

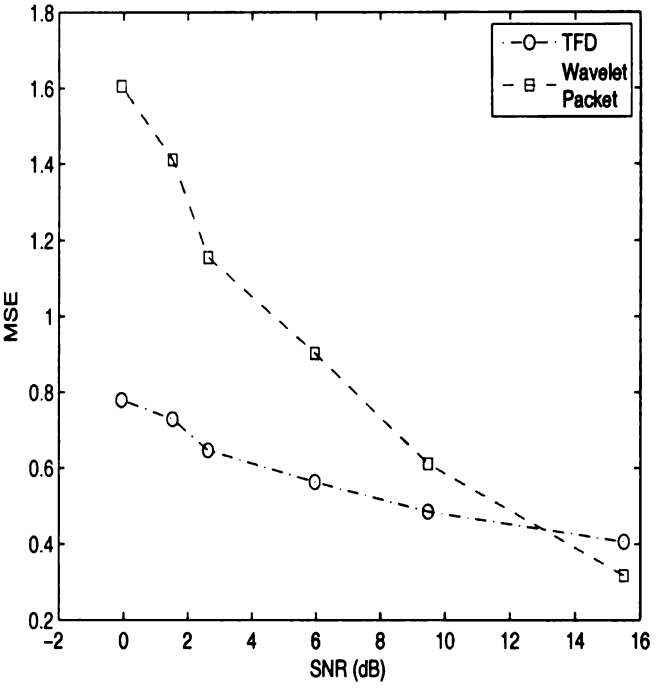


Figure 3.5. Comparison of MSE versus SNR for extracted sources with TFD and WP

CHAPTER 4

SOURCE SEPARATION RESULTS FOR ERP SIGNALS

4.1 Data

The ERP data analyzed in this thesis was recorded at Ormond and Hazel Hunt ERP Lab at the University of Michigan. The study consisted of 10 Subjects and 6 electrodes. The electrodes that were used in the recording are F3, F4, Cz, P3, P4, and Oz. The subjects are shown visual stimuli in the form of words related to their particular psychopathologoy. Two sets of words are presented to each subject, prime and target. The prime stimulus is always presented subliminally with stimulus duration less than 5ms. One second after the prime stimulus, the target stimulus is presented supraliminally, i.e. the subject is aware of the stimulus. Three groups of words are uniquely developed for each subject. One category, unpleasant word, is the same for all patients, where each word is a generally unpleasant word, such as 'cheating', 'cancer', or 'lying'. The other two categories are unique to the subject's condition. One category is the conscious conflict category, which the words were used by the patient to describe his/her condition. Examples could be from a subject who suffers from a public eating phobia, such as 'swallowing', 'cafeteria', or 'headache'. The final category is unconscious conflict words. Experts develop a list of words related to the condition. Examples for the same subject are 'massaging muscle', 'ripped apart', or 'on my back'. In this study there are two groups for each prime and target stimuli. The prime stimuli belong to either unconscious conflict word or conscious conflict word categories, whereas the target stimuli belong to conscious conflict word and unpleasant word categories. Each set of stimulus is repeated 49 times resulting in a total of 196 trials per electrode per subject. The data used in analysis was for 1s in duration upon presentation of the supraliminal stimulus, and

only the data from one subject was used.

4.2 Single-Trial ERP

The goal in single-trial ERP analysis is to be able to extract individual underlying sources in the brain which are generated in a localized area. With successful source extraction, analysis of individual responses of the brain can be performed, and the dynamic variability of the ERP responses can be compared on a trial to trial basis. In this way, observations can be made on all factors affecting subject's performance. A comparison is made between the algorithm outlined in section 3.4 and ICA as outlined applied to the same data. Both BSS techniques are applied to all 196 trials of data available.

In application of the two-stage approach, first, a number of sources to extract, k, must be chosen. This value was empirically chosen, and is chosen such that it is greater than the number of electrodes, 6. Multiple trials were run under a selection for k. If no sources extracted appeared spurious, k was incremented, since it was shown in [35] that choosing k larger than actual number of sources still results in successful extraction of sources. As k increased, sources began to show up in the results that had only spurious activity, incrementation was stopped and k was chosen. Experiments were done using 32 frequency bins, for which k was 8, and using 128 frequency bins, for which k was 14 and 16.

ICA was then applied to the same data. Since only 6 mixtures are used, ICA can only extract 6 components per trial. The results for ICA are in the time domain, so they are converted to the time-frequency plane at the different frequency resolution levels using 32 or 128 frequency bins for comparison.

Examples of single-trial results are shown in Figure 4.1, Figure 4.2, and Figure 4.3. Figure 4.1 shows results of one trial when 32 frequency bins were used in the TFD, while Figure 4.2 and Figure 4.3 each show results for one trial with 128 frequency bins.

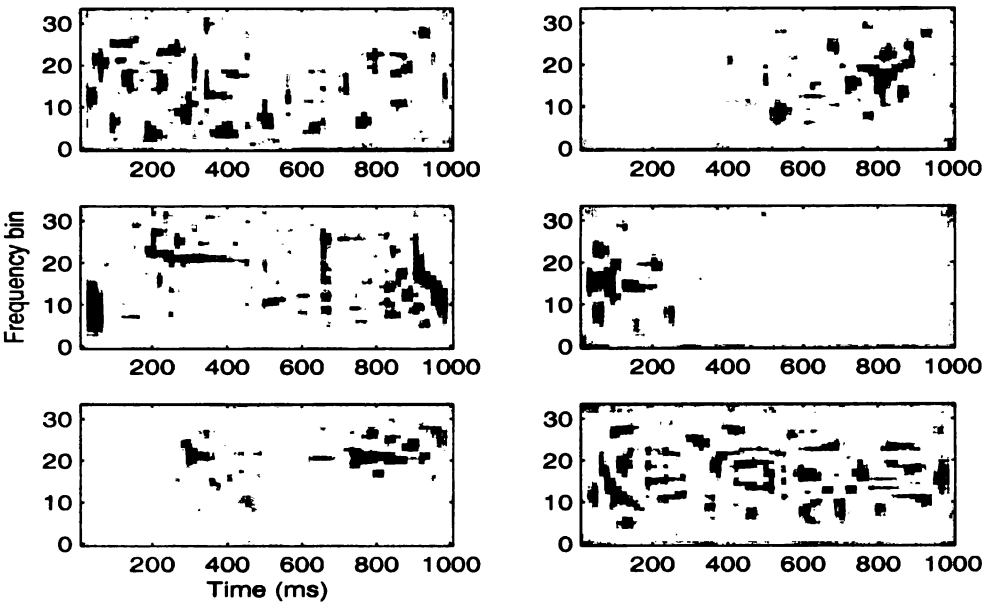
Similar results are obtained over all 196 trials. Using more frequency bins provides more data points, which provides for a more robust separation so that more sources can be extracted. This increase in data points comes at the cost of more computation time. The sources from the proposed technique show in general less activity, i.e. more sparsity, in the time-frequency plane than the sources from ICA. As more sources are extracted in the proposed technique, source representation becomes more sparse. It is, however, difficult to compare results on the single-trial level here since the underlying source generators are actually not known, and since a different number of components are extracted from each technique. It is also difficult because there are 196 individual trials to try to quantify. An attempt must be made to generalize the results.

4.3 Measures of Evaluation

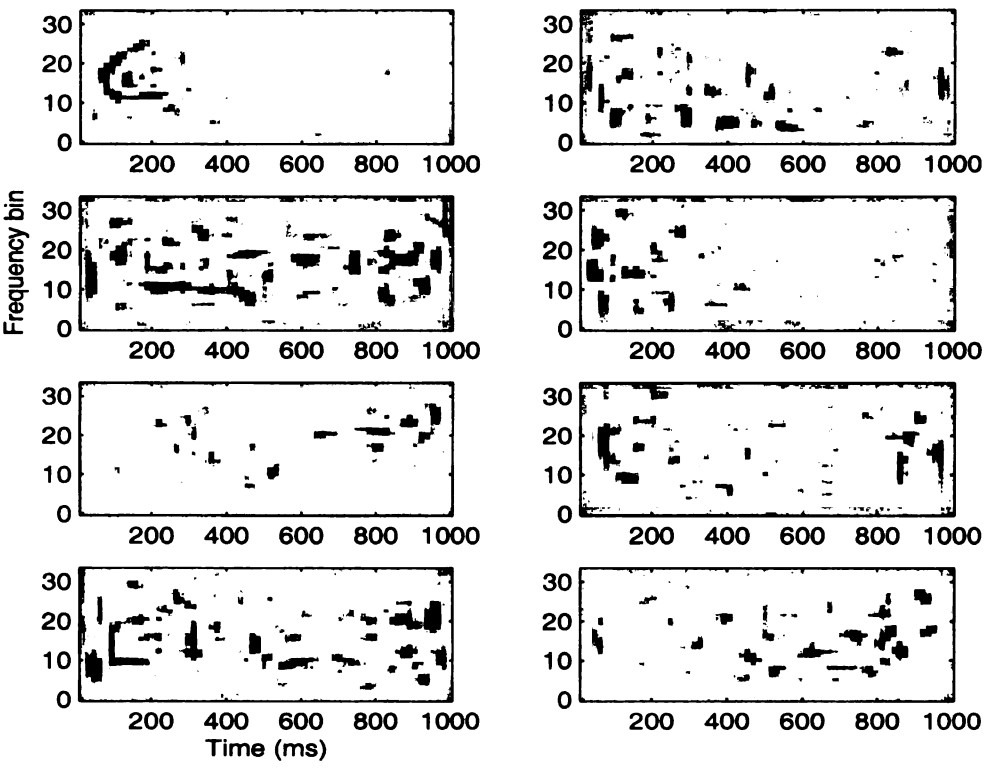
In order to evaluate the performance of ICA and the proposed UBSS method, the single-trial results are put together in their respective groups depending on stimulus type. K-means clustering is carried out over all extracted components from the subject and the extracted cluster centers represent similar components across all trials. These components are then representative of the most prevalent sources extracted throughout all the trials for each stimulus. Evaluation of these cluster centers is then carried out in attempt to quantify the general results of ICA to those of the proposed method.

4.3.1 Data Reduction

The results of one trial are represented by the matrix, $\mathbf{S}_v$, which is of size $N \times P$. Each component in time-frequency is first vectorized to form a vector of length P, which is in our case equal to either 2112 or 8256. These vectors are then put into a matrix. This represents N extracted components from trial i, each over P time-frequency points. For the data reduction of all results for a particular stimulus, the extracted matrices over all trials are each appended to form a new matrix, $\check{\mathbf{S}}_u$, such

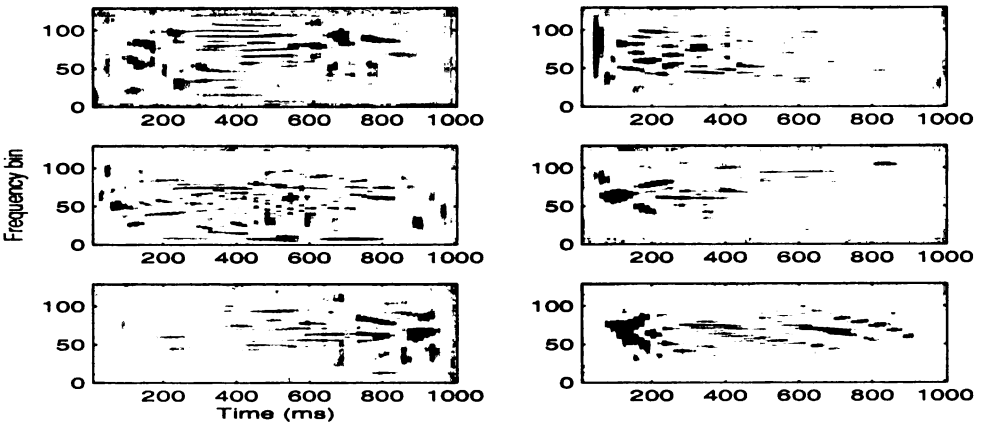


(a)

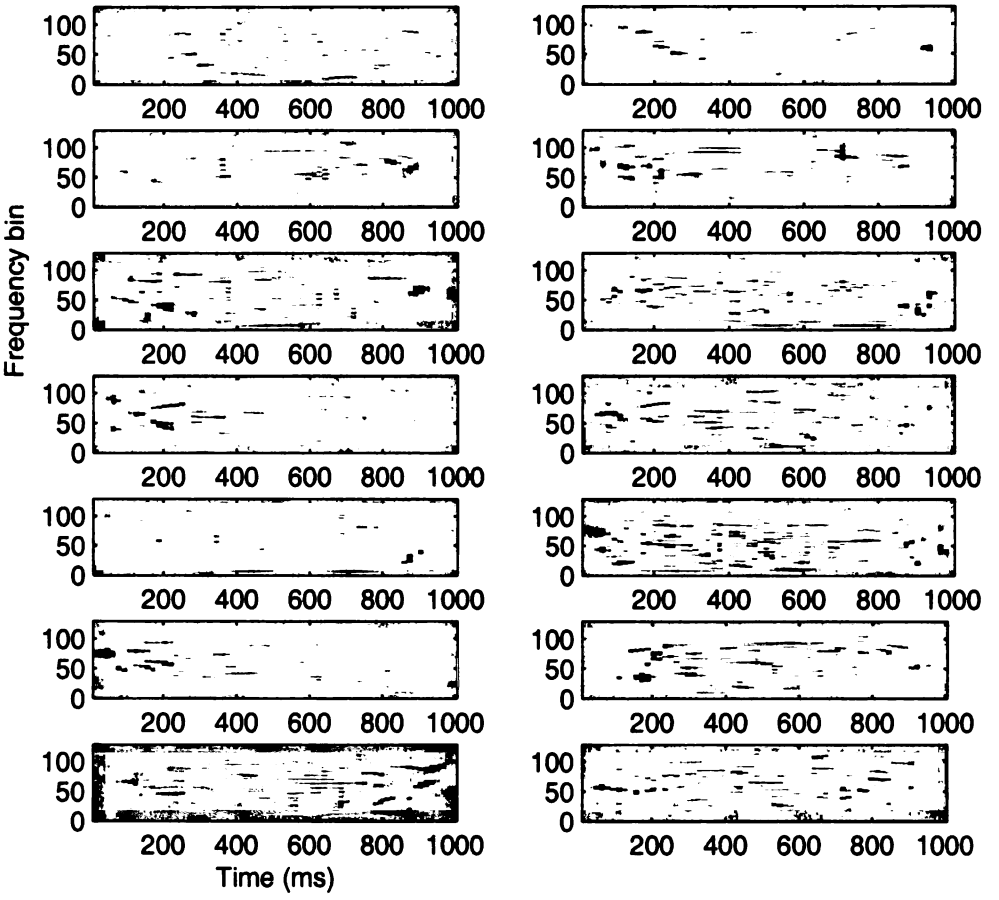


(b)

Figure 4.1. Single trial results using 32 frequency bins. (a) 6 extracted sources from ICA. (b) 8 extracted sources from the proposed method.

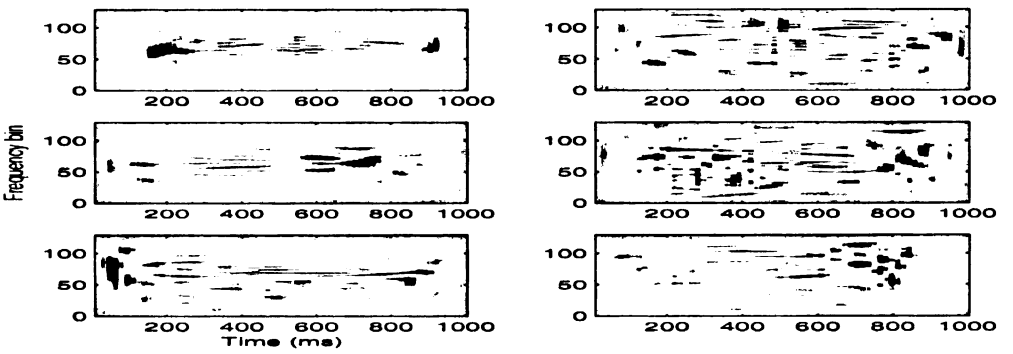


(a)

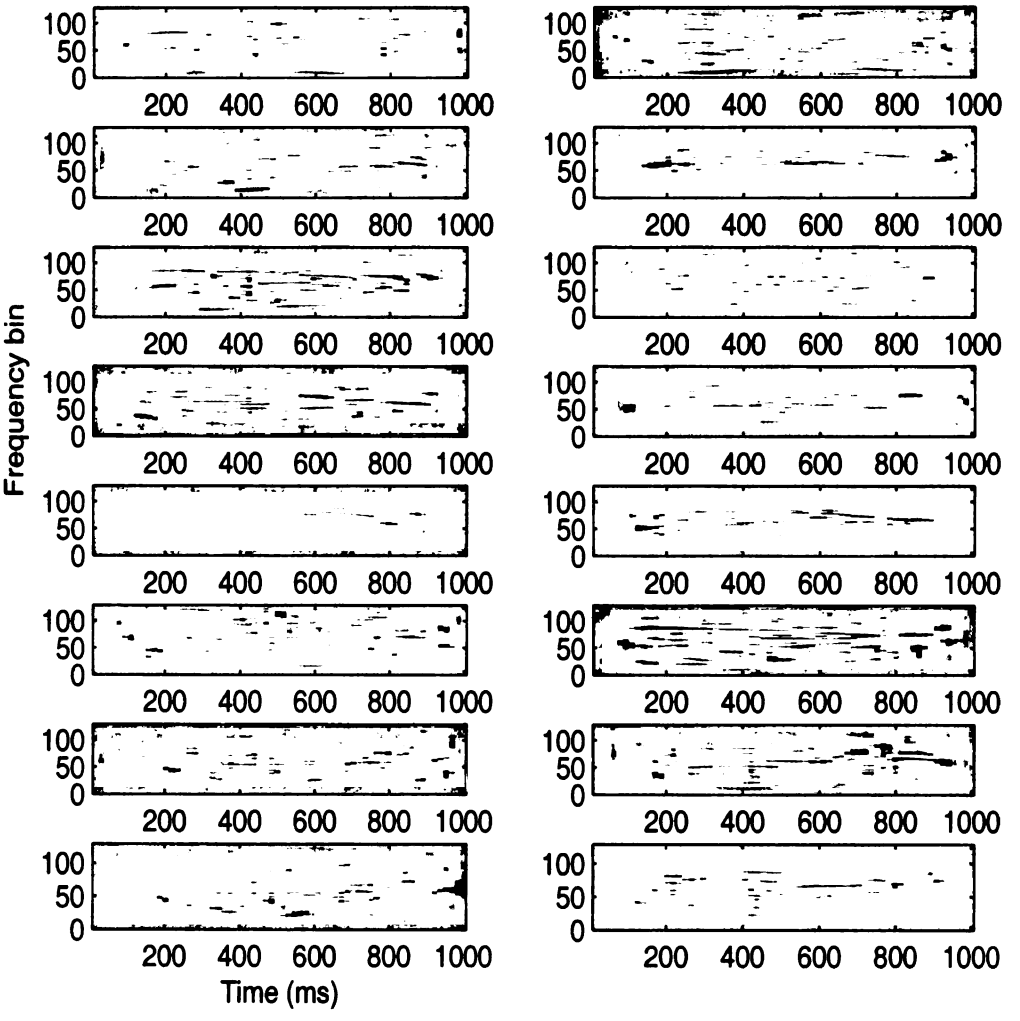


(b)

Figure 4.2. Single trial results using 128 frequency bins. (a) 6 extracted sources from ICA. (b) 14 extracted sources from the proposed method.



(a)



(b)

Figure 4.3. Single trial results using 128 frequency bins. (a) 6 extracted sources from ICA. (b) 16 extracted sources from the proposed method.

that

$$\check{\mathbf{S}}_u = \begin{bmatrix} \mathbf{S}_1 \\ \mathbf{S}_2 \\ \vdots \\ \mathbf{S}_{49} \end{bmatrix} = \begin{bmatrix} s_1^1(1) & \cdots & s_1^1(P) \\ \vdots & \vdots & \vdots \\ s_N^1(1) & \cdots & s_N^1(P) \\ s_1^2(1) & \cdots & s_1^2(P) \\ \vdots & \vdots & \vdots \\ s_N^{49}(1) & \cdots & s_N^{49}(P) \end{bmatrix}, \tag{4.1}$$

where $u = \{1, 2, 3, 4\}$ represents the stimulus pair number, and $\check{\mathbf{S}}_u$ is of size $49N \times P$. Each element, $s_i^v(j)$, is one time-frequency point of source i from trial v.

K-means clustering is then carried out on each $\check{\mathbf{S}}_u$ where each of its rows is grouped into one of K clusters based on its squared Euclidean distance to that cluster center as described in section 3.3. The clustering algorithm is run 10 times to avoid randomness in the final cluster results. We run K at 8, 10, and 12 to get an idea of how a different number of components may affect the outcome.

The rows of $\check{\mathbf{S}}_u$ are then grouped by a hierarchical clustering method based on the results of the 10 k-means runs. A matrix, $\mathbf{R}$, of zeros of size $49N \times 49N$ is created. Each entry is updated iteratively. The entry, r_{ij} represents how many times out of 10, row i of $\check{\mathbf{S}}_u$ was grouped into the same cluster as row j of $\check{\mathbf{S}}_u$. This matrix then serves as a similarity measure, the more times two sources were grouped together by k-means, the more similar they are. All diagonal entries, r_{ii}, represent how many times each source was grouped with itself. These entries are ignored because they are all 10 and are meaningless.

A hierarchical clustering is then carried out using the similarity matrix, $\mathbf{R}$, as its distance metric. In the first step, each row of $\check{\mathbf{S}}_u$ is in its own cluster. The second step then groups all rows together with a similarity value of 10 in the matrix, $\mathbf{R}$. Next, all rows with similarity of 9 are grouped. If a group already exists, then the average similarity between one row and all rows already in the cluster is used. The next step

will then group together a cluster with another cluster or individual row if it has the highest similarity value remaining. If not, then all rows with similarity value of 8 are grouped together. This is repeated until the number of clusters is reduced to K. Cluster centers are then calculated by the mean of the time-frequency components in each cluster, and these are the components that categorize all single-trial ERP results. An example of a set of extracted components is shown in Figure 4.4.

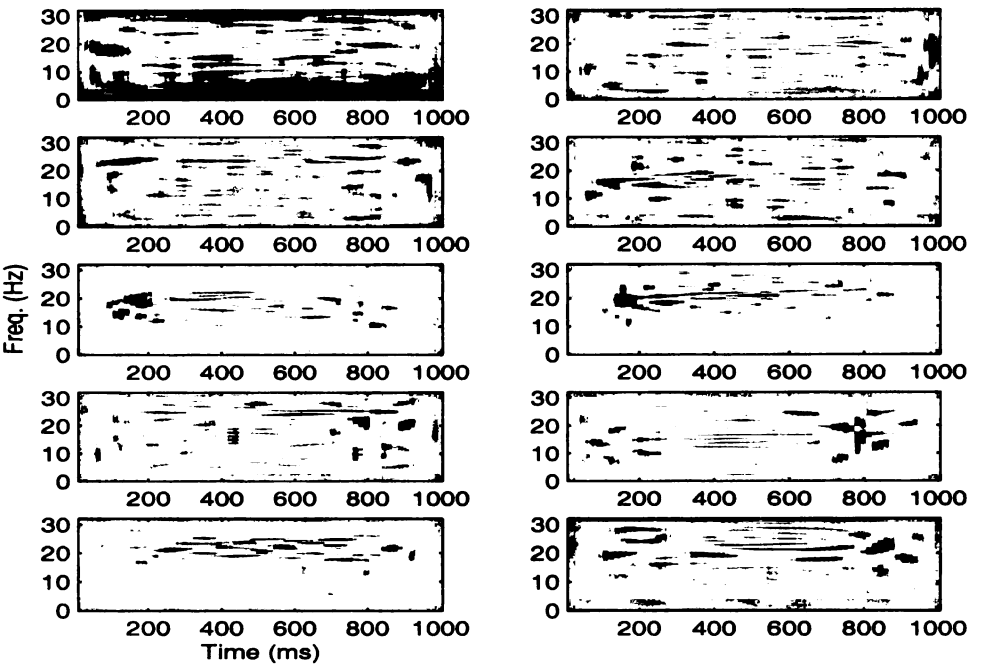
4.3.2 Data Variance Explained

After the extracted single trial components are reduced using clustering methods, a comparison of component quality is then sought out. The K extracted components from all trials of the two-stage approach must be compared with the K extracted components from all trials of ICA. It is possible that the better representation of the underlying activity will be able to explain more of the variance of the original data. The extracted clusters are represented by the $K \times P$ matrix $\mathbf{C}_u$, $u = \{1,2,3,4\}$. The space spanned by these components can be found by Gram-Schmidt orthoganalization, to produce the $K \times P$ orthogonal matrix $\hat{\mathbf{C}}_u$, $u = \{1,2,3,4\}$. These orhtogonal components are projected back onto the original electrode measurements, $\mathbf{X}_u^l$, $u = \{1,2,3,4\}$, $l = \{1,\ldots,6\}$. This is calculated by:

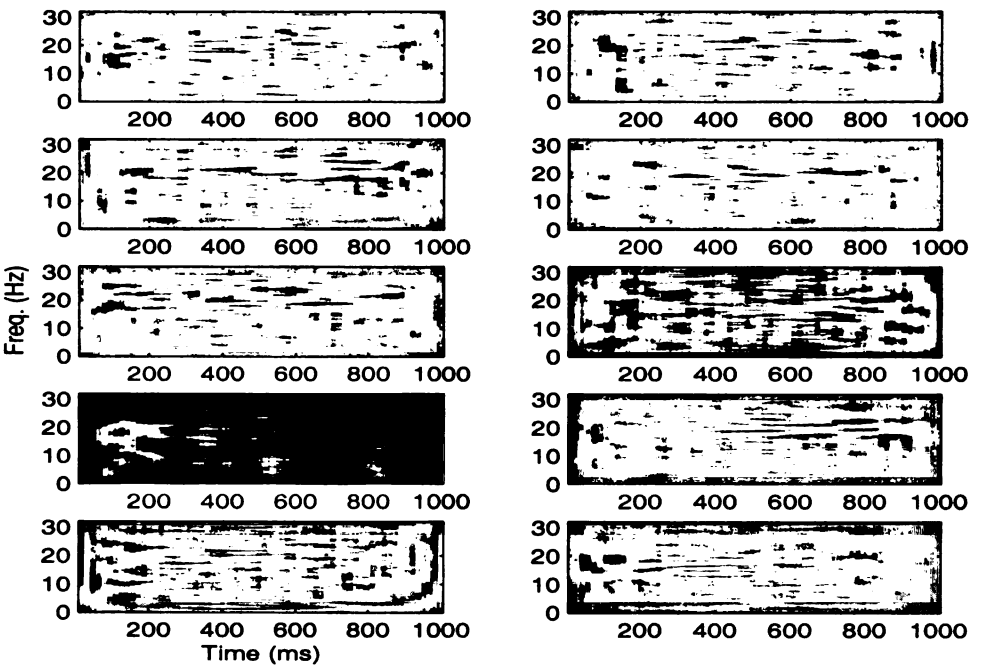
$$\mathbf{D}_u^l = (\hat{\mathbf{C}}_u \mathbf{X}_u^{l^T})^2$$

$$= \begin{bmatrix} (\sum_{j=1}^{P} \hat{c}_1^u(j) x_1^{u,l}(j))^2 & (\sum_{j=1}^{P} \hat{c}_1^u(j) x_2^{u,l}(j))^2 & \cdots & (\sum_{j=1}^{P} \hat{c}_1^u(j) x_{49}^{u,l}(j))^2 \\ (\sum_{j=1}^{P} \hat{c}_2^u(j) x_1^{u,l}(j))^2 & (\sum_{j=1}^{P} \hat{c}_2^u(j) x_2^{u,l}(j))^2 & \cdots & (\sum_{j=1}^{P} \hat{c}_2^u(j) x_{49}^{u,l}(j))^2 \\ \vdots & \vdots & \vdots & \vdots \\ (\sum_{j=1}^{P} \hat{c}_K^u(j) x_1^{u,l}(j))^2 & (\sum_{j=1}^{P} \hat{c}_K^u(j) x_2^{u,l}(j))^2 & \cdots & (\sum_{j=1}^{P} \hat{c}_K^u(j) x_{49}^{u,l}(j))^2 \end{bmatrix}, \tag{4.2}$$

where each entry, $d_{mv}^{u,l}$, is at row m and column v and represents the projection of cluster m back onto trial v of electrode l for stimulus group u. The variance of the



(a)



(b)

Figure 4.4. Results of component clustering over all single-trial results for stimulus group $u = 1$. (a) Components extracted using ICA. (b) Components extracted using the proposed method.

data, z, at electrode l explained by all extracted components from stimulus group u is then:

$$z_u = \sqrt{\frac{1}{49}\frac{1}{6}\sum_{m=1}^{K}\sum_{v=1}^{49}\sum_{l=1}^{6} d_{mv}^{u,l}}. \tag{4.3}$$

The results are shown in Table 4.1. It is seen that in general more variance is explained by the components extracted from ICA than by the two-stage approach when 32 frequency bins are used. As resolution increases, the performance of the two-stage approach improves with respect to ICA, and in almost all cases is better. This does not necessarily mean that ICA is outperforming the two-stage approach in the lower resolution, however. The two-stage approach seeks to find the sparsest sources possible. ICA is seeking to find maximally independent sources. When comparing an equal number of extracted components from the two methods, those with less sparse representations (from ICA) project better back to the original measurements, but it is likely that these components are linear sums of further reducible sources. The extracted components must also be evaluated for time-frequency localization and sparseness.

Table 4.1. Mean measure of l_1 norm to show sparsity.

Running Conditions	u=1		u=2		u=3		u=4	
	UBSS	ICA	UBSS	ICA	UBSS	ICA	UBSS	ICA
k=8 K=8	0.467	0.457	0.452	0.453	0.452	0.461	0.461	0.463
k=8 K=10	0.472	0.473	0.462	0.472	0.460	0.474	0.472	0.476
k=8 K=12	0.471	0.482	0.464	0.482	0.468	0.480	0.476	0.490
k=14 K=8	0.400	0.370	0.386	0.368	0.375	0.366	0.387	0.368
k=14 K=10	0.408	0.375	0.396	0.364	0.384	0.366	0.396	0.366
k=14 K=12	0.417	0.387	0.404	0.389	0.390	0.385	0.403	0.390
k=16 K=8	0.399	0.367	0.386	0.367	0.374	0.367	0.385	0.371
k=16 K=10	0.409	0.376	0.396	0.379	0.383	0.378	0.395	0.382
k=16 K=12	0.410	0.387	0.393	0.388	0.391	0.384	0.390	0.391

4.3.3 Measurement of Sparsity

The level to which the extracted components are sparse, disjoint, and localized in the time-frequency plane all speak to how close they may be to an actual underlying source in the brain. The components obtained from the clustering method described in the previous section are evaluated based on these factors. Sparsity will be measured using the l_1-norm, disjointness using the total inner product between the components, and localization using a measure of entropy on the time-frequency plane.

Since a sparse component must have most of its values close to zero, the l_1-norm is a good measurement of how sparse a component is and a smaller l_1-norm means a sparser component. Each row of the extracted cluster matrix, $\mathbf{C_u}$, represents one extracted component. Thus each component's sparsity is measured with

$$\sum_{m=1}^{P} |c_i^u(m)|, \tag{4.4}$$

where i represents component number.

Disjointness between two components is measured by using the inner product. A summation of all the pairwise inner products between L components represents a total level of disjointness over all extracted components. This is computed as

$$\sum_{i \neq j} \sum_{m=1}^{P} |c_i^u(m) c_j^u(m)|, \tag{4.5}$$

where u refers to stimulus group, and P is the number of time-frequency points.

Time-frequency localization of each component is computed using a measurement of entropy. This is calculated as

$$-\sum_{m=1}^{P} |c_i^u(m)| log_2 |c_i^u(m)|, \tag{4.6}$$

where u refers to stimulus group, and i is the component number, between 1 and L. A smaller entropy corresponds to a more localized component.

The results calculated for these parameters are shown in Table 4.2, Table 4.3, and Table 4.4. This shows that under the two-stage approach, the extracted components are typically more sparse, localized, and disjoint than the extracted components under ICA. This means that under the two-stage approach, the components are more likely a closer representation of a true source. As time-frequency resolution increases, the extracted components from the two-stage approach represent more of the original data variance, while still remaining sparse, localized and disjoint.

Table 4.2. Mean measure of l_1 norm to show sparsity.

Running Conditions	u=1		u=2		u=3		u=4	
	UBSS	ICA	UBSS	ICA	UBSS	ICA	UBSS	ICA
k=8 K=8	29.93	35.12	34.85	33.12	29.43	34.80	30.07	35.20
k=8 K=10	29.74	33.95	29.40	33.26	29.87	34.29	30.87	34.25
k=8 K=12	30.36	33.44	29.57	33.11	30.19	33.73	31.40	33.68
k=14 K=8	55.84	67.04	55.65	66.67	55.81	68.00	56.34	67.67
k=14 K=10	56.96	62.03	55.48	65.61	51.39	66.80	57.64	66.27
k=14 K=12	58.23	64.57	55.72	64.04	54.41	62.87	56.90	64.92
k=16 K=8	56.53	67.31	55.08	66.51	56.16	68.16	51.84	66.97
k=16 K=10	56.67	65.64	54.72	65.14	54.73	66.92	53.40	66.69
k=16 K=12	52.51	65.24	51.17	65.07	55.42	64.16	55.32	65.78

Table 4.3. Mean measure of entropy to show time-frequency localization.

Running Conditions	u=1		u=2		u=3		u=4	
	UBSS	ICA	UBSS	ICA	UBSS	ICA	UBSS	ICA
k=8 K=8	162.75	193.89	190.27	181.08	159.95	191.92	163.28	194.29
k=8 K=10	161.36	186.94	159.47	183.61	162.21	187.11	167.49	188.98
k=8 K=12	164.74	184.37	160.44	182.31	164.03	185.73	170.25	185.66
k=14 K=8	360.03	433.63	358.35	430.80	354.03	438.95	362.47	437.52
k=14 K=10	366.53	396.14	356.92	423.49	319.68	430.92	370.22	428.26
k=14 K=12	374.67	417.32	358.58	413.65	339.23	401.85	366.04	418.42
k=16 K=8	364.45	435.27	355.34	429.95	361.31	439.87	329.96	433.01
k=16 K=10	365.20	424.30	352.87	421.00	352.12	432.04	340.10	430.84
k=16 K=12	331.54	421.19	325.11	418.67	356.82	411.23	356.36	424.66

Table 4.4. Measure of disjointness by correlation between components.

Running Conditions	u=1		u=2		u=3		u=4	
	UBSS	ICA	UBSS	ICA	UBSS	ICA	UBSS	ICA
k=8 K=8	4.10	4.73	5.63	5.85	2.96	3.56	3.00	3.68
k=8 K=10	6.27	6.85	4.07	4.65	4.09	4.47	4.15	4.59
k=8 K=12	8.94	9.58	6.55	7.17	6.48	6.96	6.48	7.13
k=14 K=8	2.84	3.46	2.26	2.92	3.22	3.78	2.27	2.85
k=14 K=10	4.21	4.27	3.61	4.12	3.44	4.00	3.63	4.24
k=14 K=12	6.10	6.63	5.25	5.70	4.00	4.24	5.30	5.80
k=16 K=8	2.76	3.44	2.58	3.23	2.81	3.36	2.49	3.12
k=16 K=10	4.20	4.77	3.91	4.38	4.47	4.98	3.85	4.38
k=16 K=12	5.33	5.87	4.64	5.19	5.43	5.92	5.68	6.17

CHAPTER 5

CONCLUSIONS AND FUTURE WORK

This thesis discusses the study of ERPs using EEG measurements to help understand mental processes. A number of components have been identified in the time-domain using the traditional method of averaging over multiple trials of ERPs and detecting characteristic peaks. These components have been linked to general brain processes, but lack specificity because their representation is probably the summation of multiple different sources. Additionally, averaging over trials ignores trial-to-trial variability and cannot detect any changes of state in the subject. For this reason, single-trial analysis by means of BSS has become a highly researched area in its application to ERP data. Different methods of BSS from the literature are introduced. It is proposed to use a UBSS algorithm and apply it to ERP data using TFDs. EEG signals have been shown to be non-stationary and this approach does not assume stationarity unlike some other techniques. The proposed approach is also capable of extracting more sources than sensors, where many techniques cannot do so. This is important since the number of sources are unknown, and since many EEG setups do not have large electrode arrays. This UBSS approach is compared to the popular ICA algorithm when applied to the same multiple trial ERP data set. Data reduction by clustering is performed over all single-trial results to extract components that represent the results. These components were not shown to explain more variance of the original data using 32 frequency bins, but performance improved with resolution. The components were consistently more sparse using the proposed UBSS technique than with ICA, showing that ICA probably tends to extract components that are sums of sources, and can help explain the higher correlation value to the original data. The UBSS technique provided components that are more localized in the time-

frequency plane and that are more distinct from each other than did ICA. Because the components extracted using the UBSS technique had these desired properties and explained more of the data variance using 128 frequency bins, the method presented in this thesis can provide a more useful means of data reduction than using PCA or ICA. Even in the case of lower resolution, the variance explained is comparable between the two methods, while using the UBSS algorithm provides the desired qualities of the extracted components.

It would be beneficial to be able to run this UBSS algorithm more efficiently, since using TFDs increases the amount of data to examine by so much. Being able to use higher resolution TFDs will provide more reliable results. Another problem with this method is the arbitrary selection of how many sources to extract. It would be more efficient to have the algorithm automatically select this number. The requirement that the sources must be approximately disjoint limits the algorithm. If this assumption could be relaxed, results could be more reliable since neuronal sources may not be disjoint. Further study should be done to compare extracted sources to available cortical data or the technique used with source localization techniques for better understanding of how the components are generated.

BIBLIOGRAPHY

BIBLIOGRAPHY

[1] H. A. Jasper, "The ten-twenty system of the International Federation," Electroencephalography and Clinical Neurophysiology, vol. 10, pp. 371-375, 1958.

[2] R. Conduit, "Polysomnographic (PSG) Recording in Humans," http://www.adinstruments.com/downloads/contributions/PolysomnographicRecording.pdf?template=education.

[3] J. W. Rohrbaugh, R. Parasuraman, and R. J. Johnson Jr., "Event-Related Brain Potentials," Oxford University Press, New York, 1990.

[4] J. S. Barlow, "The Electroencephalogram," The MIT Press, Cambridge, 1993.

[5] A.S. Gevins, A. Remond, "Handbook of Electroencephalography and Clinical Neurophysiology Volume 1," Elsevier, New York, 1987.

[6] M. Modarreszadeh, R. N. Schmidt, "Wireless, 32-Channel, EEG and Epilepsy Monitoring System," Proc. 19th International Conference - IEEE/EMBS, pp. 1157-1160, 1997.

[7] K. Osaka, S. Chiba, T. Tanioka, C. Kawanishi, I Nagamine, F. Ren, S. Kuroiwa, T. Tada, R. Yamashita, M. Kishimoto, M. Nishimura, A. Yamamoto, R. C. Locsin, Y. Takasaka, "Electroencephalographic Activities and its Clinical Application," Proc. of NLP-KE, 2005.

[8] J. R. Wolpaw, N. Birbaumer, D. J. McFarland, G Pfurtscheller, T. M. Vaughan, "Brain-Computer Interfaces for Commication and Control," Clinical Neurophysiology, vol. 113, pp. 767-791, 2002.

[9] M. Fabiani, G. Gratton, M. G. H. Coles, "Event-Related Brain Potentials," Chapter 3 in "Handbook of Psychophysiology," Cambridge, 2002.

[10] A. Isaksson, A. Wennberg, L. H. Zetterberg, "Computer Analysis of EEG Signals with Parametric Models," Proc. of the IEEE, vol. 69 iss. 4, pp. 451-461, 1981.

[11] V. Goel, A. M. Brambrink, A. Baykal, R. C. Koehler, D. F. Hanley, N. V. Thakor, "Dominant Frequency Analysis of EEG Reveals Brain's Response During Injury and Recovery," IEEE Tran. on Biomedical Engineering, pp. 1083-1092, 1996.

[12] M. T. Tebano, M. Cameroni, G. Gallozzi, A. Loizzo, G. Palazzino, G. Pezzini, and G. F. Ricci, "EEG Spectral Analysis After Minor Head Injury in Man," Electroenceph. Clin. Neurophysiol., vol. 70, pp. 185-189, 1988.

[13] E. Schwarz, P. Kielholz, V. Hobi, L. Goldberg, U. Gilsdorf, M. Hofstetter, D. Ladewig, P. C. Miest, G. Reggiani, R. Richter, "Alcohol-Induced Biphasic Background and Stimulus-Elicited EEG Changes in Relation to Blood Alcohol Levels," Int. J. Clin. Pharmacol. Ther. Toxicol., vol. 19 iss. 3, pp. 102-111, 1981.

[14] T.P. Jung, S. Makeig, M. Westerfield, J. Townsend, E. Courchesne, T.J. Sejnowski, "Analysis and Visualization of Single-Trial Event-Related Potentials," Human Brain Mapping, vol. 14, pp. 166-185, 2001.

[15] O. Faugeras, F. Clement, R. Deriche, R. Keriven, T. Papadopoulo, J. Roberts, T. Vieville, F. Devernay, J. Gomes, G. Hermosillo, P. Kornprobst, and D. Lingrand, "The inverse EEG and MEG problems: The adjoint state approach I: The continuous case," INRIA Research Report 3673, June 1999.

[16] L. Cohen, "Time-Frequency Analysis," Prentice Hall, New Jersey, 1995.

[17] E. M. Bernat, W. J. Williams, and W. J. Gehring, "Decomposing ERP time-frequency energy using PCA," Clinical Neurophysiology, vol. 116, pp. 13141334, 2005.

[18] W. J. Williams, "Reduced Interference Distributions: Biological Applications and Interpretations," Proc. of the IEEE, vol. 84, iss. 9, pp. 1264-1280, 1996.

[19] S. Mallat and Z. Zhang, "Matching Pursuits With Time-Frequency Dictionaries," IEEE Trans. on Signal Processing, vol. 41, pp. 3397-3415, 1993.

[20] P. J. Durka, D. Ircha, and K. J. Blinowska, "Stochastic Time-Frequency Dictionaries for Matching Pursuit," IEEE Trans. on Signal Processing, vol. 49 no. 3, pp. 507-510, 2001.

[21] T. Demiralp, J. Yordanova, V. Kolev, A. Ademoglu, M. Devrim, and V.J Savar, "Time-Frequency Analysis of Single Sweep Event-Related Potentials by Means of Fast Wavelet Transform," Brain and Language, vol. 66, pp.129-145, 1999.

[22] A. Cichocki and S. Amari, "Adaptive Blind Signal and Image Processing: Learning Algorithms and Applications," Wiley, New York, 2002.

[23] A. J. Bell and T. J. Sejnowski, "An Information-Maximization Approach to Blind Separation and Blind Deconvolution," Neural Comput, vol. 7, pp. 1129-1159, 1995.

[24] P. Comon, "Independent Component Analysis, A New Concept?" Signal Processing, vol. 36, pp. 287-314, 1994.

[25] A Delorme and S. Makeig, "EEGLAB: An Open Source Toolbox for Analysis of Single-trial EEG Dynamics Including Independent Component Analysis," Jounal of Neuroscience Methods, vol. 134, pp. 9-21, 2004.

[26] S. Aviyente, E. Bernat, S. Malone and W. Iacono, Analysis of Event-Related Potentials Using PCA and Matching Pursuit on the Time-Frequency Plane, IEEE International Conference of the Engineering in Medicine and Biology Society, pp. 2454-2457, 2006.

[27] A. Delorme, S. Makeig, "EEGLAB: An open source toolbox for analysis of single-trial EEG dynamics including independent component analysis," Journal of Neuroscience Methods, vol. 134, pp. 9-21, 2004.

[28] O. Yilmaz and S. Rickard, "Blind Separation of Speech Mixtures via Time-Frequency Masking," IEEE Transactions on Signal Processing, vol. 52 no. 7, pp. 1830-1847, 2004.

[29] A. Belouchrani and M.G. Amin, "Blind Source Separation based on Time-Frequency Signal Representation," IEEE Trans. on Signal Processing, vol. 46, pp. 2888-2897, 1998.

[30] A. Belouchrani, K. Abed-Meraim, M.G. Amin, and A.M. Zoubir, "Blind Separation of Nonstationary Sources," IEEE Signal Processing Letters, vol. 11 no. 7, 2004.

[31] M.G. Amin and Y. Zhang, "Signal Averaging of Time-Frequency Distributions for Signal Recovery in Uniform Linear Arrays," IEEE Trans. on Signal Processing, vol. 48 no. 10, pp. 2892-2902, 2000.

[32] N. Linh-Trung, A. Belouchrani, K. Abed-Meraim, and B. Boashash, "Separating More Sources than Sensors Using Time-Frequency Distributions," Proc. Int. Symp. on Signal Processing and its Applications, pp. 583-586, 2001.

[33] P. Bofill, M. Zibulevsky, "Underdetermined Blind Source Separation Using Sparse Representations," Signal Processing, vol. 81, pp. 2353-2362, 2001.

[34] Y. Li, A. Cichocki, and S. Amari, "Blind Estimation of Channel Parameters and Source Components for EEG Signals: A Sparse Factorization Approach," IEEE Trans. on Neural Networks, vol. 17 no. 2, 2006.

[35] Y. Li, A. Cichocki, S. Amari, "Sparse Component Analysis for Blind Source Separation With Less Sensors Than Sources," 4th Int. Symp. on ICA and BSS, 2003.

[36] I. F. Gorodnitsky and B. D. Rao, "Sparse Signal Reconstruction from Limited Data Using FOCUSS: A re-weighted Minimum Norm Algorithm," IEEE Trans. on Signal Processing, vol. 45, pp. 600-616, 1997.

[37] D. L. Donoho and M. Elad, "Maximal Sparsity Representation via l_1 Minimization," Proc. Nat. Acad. Sci., pp. 2197-2202, 2003.

[38] Z. Shan, J. Swary, S. Aviyente, "Underdetermined Source Separation in the Time-Frequency Domain," IEEE Conference on Acoustics, Speech, and Signal Processing, vol. 3, pp.945-948, 2007.

[39] R. O. Duda, P. E. Hart, and D. G. Stork, "Pattern Classification," Wiley, New York, 2001.

[40] S. S. Chen, D. L. Donoho, and M. A. Saunders, "Atomic Decomposition by Basis Pursuit," SIAM J. Scientific Comp., vol. 20, pp. 33-61, 1999.


MICHIGAN STATE UNIVERSITY LI
3 1293 02956 2