UNIVERSITY OF SOUTHAMPTON

FACULTY OF ENGINEERING, SCIENCE AND MATHEMATICS

School of Electronics and Computer Science

Digital Signal Processing Techniques for Detection Applied to Biomedical Data

by

Hubert Dietl

A doctoral thesis submitted in partial fulfilment of the requirements for the award of Doctor of Philosophy

May 2005

UNIVERSITY OF SOUTHAMPTON

ABSTRACT

FACULTY OF ENGINEERING, SCIENCE AND MATHEMATICS

SCHOOL OF ELECTRONICS AND COMPUTER SCIENCE

Doctor of Philosophy

Digital Signal Processing Techniques for Detection Applied to Biomedical Data

by Hubert Dietl

This thesis is concerned with the application of digital signal processing methods to different kinds of biomedical data by extracting their features and classifying them.

In more detail we aim at a detection by feature extraction, feature selection and classification and apply these methods to two different kinds of data whereby the type of biomedical data studied is restricted to stimulus response electro-physiological data.

Firstly, we give a review and a definition of linear transformations that can be employed for the analysis, parameterisation or compression of biomedical data. We are particularly considering the wavelet transform, the wavelet packet transformation and the Gabor expansion under the aspect of data defined on a finite interval. For this, we introduce a novel matrix notation for each transformation method. Also, appropriate signal extension methods are described for data on finite intervals.

Secondly, methods for the feature selection are studied and developed. A simple energy reduction approach is stated to start with. Then, statistical tests are explained that can be used to increase the significance when only few data points are available. These methods select certain time-frequency coefficients and the separability performance of each selected coefficient can be evaluated by a receiver operating characteristic (ROC) analysis. The ROC analysis is used to develop a signal-to-noise-like criterion, that selects and combines significant time-frequency coefficients to a coefficient set for which a separability can be stated. Also, the found coefficient set can be again evaluated by ROC analysis.

Thirdly, the classification method is introduced by support vector machines (SVM) starting with an introduction to learning theory, followed by the SVM theory. Then, we show how SVM can be used for detection of biomedical signals by introducing a connection to a diagnostic test. Also, multi-class SVM classifiers are stated with the novelty of introducing a neutral class. Moreover, it is shown that the non-linear decision boundary found by the SVM can also be evaluated by a ROC analysis.

The first application of some of the introduced signal processing tools comprises data from subjects that suffer from panic disorder. The feature selection is shown for statistical tests based on time-frequency transformed data. This approach is confirmed by the use of SVM where better separability results are obtained for the parameterised data than for the unparameterised data.

The second application is the development of a differential diagnosis method for determining cochlear hearing loss based on time-frequency transformed otoacoustic emissions. By our feature selection method a set of distinctive coefficients are determined which generalises and enhances previous studies. Then, SVM are applied for the classification which are again evaluated by a ROC analysis.

Acknowledgements

My first gratitude goes to my supervisor Dr. Stephan Weiss for his encouragements, guidances and uncountable fruitful discussions without which this thesis would not have been possible.

I am also grateful to all my colleagues in the Communications Group, both past and present, for all their support and help throughout this project, especially my office colleagues, namely Nurul Nadia Ahmad, Viktor Bale, Jin Yee Chung, Dr. Seung-Hoon Hwang, Dr. Wei Liu, Dr. Soon Xin Ng, Noor Shamsiah Othman, Ahmad Kamsani Samingan, Hafizal Mohamad, Jin Wang, Robert Maunder, Anh Pham Quang, Seung Won, Thanh Nguyen Dang, Ronald Tee and Andreas Wolfgang for the warm and friendly atmosphere in the office.

Finally, I would like to thank my family in Germany for their continual support during my research.

List of Publications

1. **H. Dietl, S. Weiss and U. Hoppe**, "Comparison of Transformation Methods to Determine Frequency-Specific Cochlear Hearing Loss Based on TEOAE", in *Proceedings of IEE Colloquium on Medical Applications of Signal Processing*, London, pp. 16/1-16/6, October 2002.

2. **H. Dietl, S. Weiss and P. Pauli**, "Time-Frequency Transform Based Panic Disorder Classification", in *Proceedings of the 2nd IEEE EMBSS Postgraduate Conference*, Birmingham, pp. 17-18, July 2003.

3. **H. Dietl and S. Weiss**, "Categorisation of Panic Disorder by Time-Frequency Methods", in *Proceedings of 37th Asilomar Conference on Signals, Systems, and Computers*, Pacific Grove, CA, November 2003.

4. **H. Dietl and S. Weiss**, "Difference Evaluation Method for Differential Diagnosis of Frequency-Specific Hearing Loss", in *Proceedings of IPEM Meeting on Signal Processing Applications in Clinical Neurophysiology*, York, February 2004.

5. **H. Dietl and S. Weiss**, "Parameterisation of Transient Evoked Otoacoustic Emissions", in *Proceedings of BIOSIGNAL 2004 International EURASIP Conference*, Brno, June 2004.

6. **H. Dietl and S. Weiss**, "Cochlear Hearing Loss Detection System Based on Transient Evoked Otoacoustic Emissions", in *Proceedings of the 3rd IEEE EMBSS Postgraduate Conference*, Southampton, pp. 21-22, August 2004.

7. **H. Dietl and S. Weiss**, "Parameterisation Comparison for the Detection of Panic Disorder Using Time-Frequency Transforms and Support Vector Machines", in *Proceedings of the 2nd International Conference on Advances in Medical Signal and Information Processing, MEDSIP*, Malta, September 2004.

8. **H. Dietl and S. Weiss**, "Detection of Cochlear Hearing Loss Applying Wavelet Packets and Support Vector Machines", in *Proceedings of 38th Asilomar Conference on Signals, Systems, and Computers*, Pacific Grove, CA, November 2004.

9. **H. Dietl and S. Weiss**, "A Novel Approach to Detect Cochlear Hearing Loss", in *Proceedings of 6th International Conference on Mathematics in Signal Processing*, Royal Agricultural College, Cirencester, December 2004.

Contents

Chapter 1

Introduction

This introductory chapter states firstly the underlying motivation for this contribution. Secondly, a summary over the novel methods that were developed for this thesis is given. The final part shows an overview of the following chapters.

1.1 Research Motivation

Researchers investigating biomedical data are faced with various difficulties. The signal-to-noise ratio may be very low, the features of the signal that are to be discovered may change over time and/or not appear at all. With this thesis a contribution is made to analyse and understand the behaviour of biomedical data better via means of digital signal processing tools. In more detail, we address the following question: How can we separate and distinguish different biomedical signals or patterns? To answer this question, we describe a parameterisation or compression of the data and how we assess or select certain of these parameters that allow us to conduct a separation and classification of the data.

In current research studies, the reason for trying to identify differences in biomedical data is to be able to detect abnormalities or diseases, [1], [2]. The application of suitable signal processing tools shall enable an automatic detection of these abnormalities or diseases, [3], [4], [5], [6], [7]. For example, [8] aims to automatically detect epileptic activity. There, the data is preliminary filtered, then parameterised, after that features are selected, and finally a neural network and a so-called expert system are used to achieve significant detection rates.

In this thesis, different approaches to separate biomedical data are shown whereby the type of biomedical data studied is restricted to stimulus response electro-physiological data. The applied approaches consist of three main steps: Firstly, the data is parameterised by time-frequency transforms. Secondly, based on the transformed data, features are to be selected which contain the differences between two data sets. Thirdly, the selected features are fed into a classifier, that yields a detection decision. Compared to [8], we aim at improving the parameterisation and feature selection for the different kinds of data analysed. Also, we use a different method for the classification than in [8]. Figure 1.1 gives a general overview of our separation and detection approach.

From biomedical data, the features are extracted by time-frequency (TF) transforms. Among the obtained TF coefficients, the ones containing the features are selected next. Finally, a classification

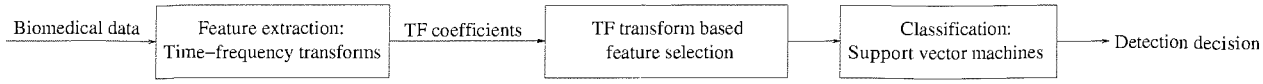

Figure 1.1: Overview of detection system.

is conducted using support vector machines (SVM),[9], yielding a binary detection decision. We will also show the development of a neutral classification result to account for data that is regarded as too difficult to be uniquely classified.

For the parameterisation of biomedical data, well-know transformations like the discrete Fourier transform or the discrete wavelet transform are commonly used, see e.g. [10], [11]. Recently, data analysis incorporating wavelet packet transforms, Gabor transforms and Karhunen-Loeve transforms have attracted attention of biomedical researchers [10],[11],[12]. In Figure 1.2 an overview over these transforms is given.

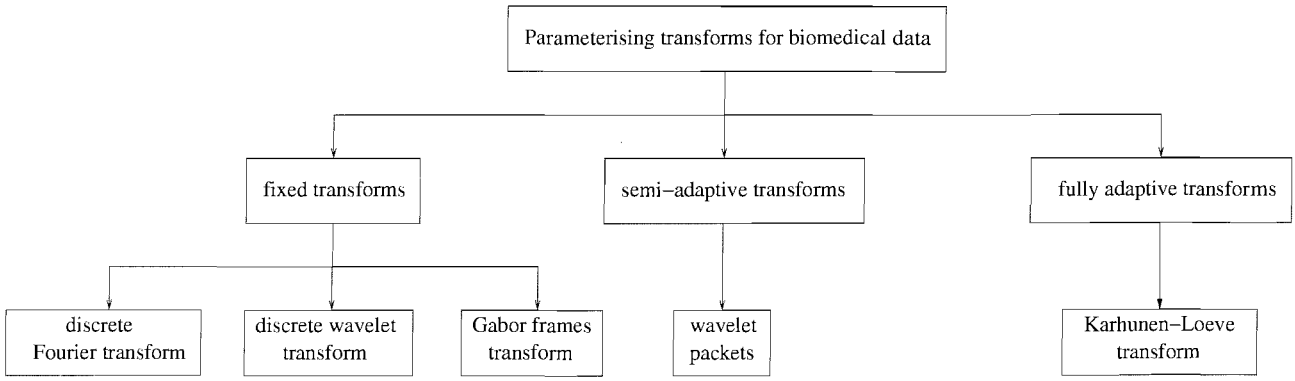

Figure 1.2: Overview of commonly used parameterisation transformations for biomedical data.

The figure illustrates the differences of the mentioned transforms with respect to their time-frequency (TF) characteristics. While the discrete Fourier transform (DFT), the discrete wavelet transform (DWT) and the Gabor frames transform (GF) possess a fixed TF tiling, the wavelet packets (WP) can be adapted to the data resulting in a specific TF decomposition. Karhunen-Loeve transforms (KLT) are directly connected to singular value decomposition (SVD), where the transform is fully adapted to the data. The short time Fourier transform (STFT) is not included in the Figure 1.2, as the GF can be regarded as a special case of the STFT, where the prototype filter defining the time window must fulfil more restricting conditions than for a STFT. There are very limited choices for the fixed transforms; the DWT depends strongly on the selected mother wavelet, the same accounts for the GF, which is dependent on the chosen prototype filter.

The second part of this work, namely the feature selection that contains the differences for data sets that ought to be distinguished is once based on a signal-to-noise(SNR)-like criterion. This criterion is used e.g. to determine cochlear status with otoacoustic emissions [13],[2]. Also, statistical tests form the basis of a feature extraction. Moreover, a simple energy reduction method is investigated.

We apply our separation methods to two different types of data: Firstly, electroencephalogram (EEG) data from subjects who suffer from panic disorder are studied. For this kind of data, only statistical analyses have been conducted so far [14]. Hence, we are interested in investigating the

performance of a separation method that is based on TF transforms and implements statistical tests for feature selection. SVM are only used for the confirmation of the feature selection and not for classification, as there was not enough data available. The results of this analysis can be used to determine the success of a therapy.

Secondly, the above mentioned otoacoustic emissions are analysed. Here, we aim at being able to distinguish and detect three types of cochlear hearing loss. We apply the DWT, WP and GF for parameterisation. The feature selection is conducted by an energy reduction method and a SNR-like criterion. SVM are used for classification.

Moreover, the introduced methods can also be applied to separate data other than presented in this thesis. E.g. within the EEG, auditory evoked potentials can be used to determine objective audiograms [5]. Our methods can further contribute to research in that field. Also, one can think of and analysing sleep disorder, which is currently studied in [15].

In the next section, on overview of our contributions to the subject of separating and classifying biomedical data is stated.

1.2 Novel Contributions

In our contributions so far, we have focused on the application of the DWT, WP and GF and have not mentioned the STFT for parameterisation because as stated above, the GF can be viewed as a STFT where the windows are time-shifted and based on low-pass filters that are linearly independent when only the positive spectrum is considered. The KLT was assumed as being too data adaptive leading to a parameterisation including too much noise. This assumption was confirmed when studying panic disorder data justifying our selection of parameterising transforms. For our second application, the WP already showed a slight adaption to the data used for adjustment and hence, the KLT was not taken into consideration, as it was expected that the results would again contain too much noise and be therefore not generally valid, as it is the case for panic disorder data.

In the following, our contributions concerning firstly the developed methodology and secondly the specific applications are described.

As stated above, the aim of this thesis is to develop a detection system with feature extraction, feature selection and classification. Each of these components consists of state-of-the-art methods, e.g. Wavelet transforms for the parameterisation and support vector machines for the classification. The combination of these methods to one new detection system represents the basis of our novel contributions. Next, a more detailed overview of the specific contributions is given.

For the parameterisation based on the TF transforms, a novel unified matrix notation is introduced. For finite discrete data, the transforms are conducted by a multiplication with a respective transformation matrix. Also, the signal extension which is an issue when dealing with finite data is incorporated into the transformation matrix. In terms of computational effort, fast implementations of the respective transforms avoiding a matrix representation may not be more efficient. After a suitable feature selection, the determination of a limited number of significant elements can be conducted faster by calculating dot products of the finite discrete data to be transformed with the corresponding

rows of the transformation matrix.

In terms of the classification, the application of SVM leads to a hard decision, meaning that data to be tested is classified to one certain class. However for electro-physiological data we studied as well as general biomedical data, it is often better not to make a decision, e.g. if a certain robustness of the classification is desired. To account for this, a neutral class is introduced. This means we establish a class for which no decision is made by the SVM classification aiming at a more reliable classifier.

Now, we continue with a description how these methods were applied to real data.

Concerning panic disorder, anxiety subjects that are presented with neutral and panic disorder triggering stimuli show different event-related brain potentials (ERP) within the EEG. We have investigated this difference by applying TF revealing transforms (DWT,WP,GF) leading to an identification of a small number of significant parameterising coefficients able to differentiate between the presented stimulus categories. The features were selected by statistical tests only. The parameterisation results in [16] were improved by incorporation further statistical tests, also leading to an improved separability in [17]. SVM are used in [18] to confirm the application of TF transforms, which yield better separability results than unparameterised data.

The second type of data we studied are otoacoustic emissions. Here, we aim to determine frequency-specific cochlear hearing loss (HL) by means of transient evoked otoacoustic emissions (TEOAE). The differentiation of three groups of frequency-specific HL is performed by parametrisation of the time-domain TEOAE responses based on three transforms DWT, WP and GF. Using an SNR-like criterion, the various transforms are tested for their ability to differentiate between the three groups of hearing ability. SVM are used for classification, including a comparison of an introduced neutral class to the case without a neutral class. Our procedure is evaluated on a large group of data from subjects. To confirm the findings for each transform method, the results are checked against the data of a second control group and can be comprehensively found in [19],[20],[21],[22],[23],[24].

The conclusion of this chapter follows next with an overview about the structure of this thesis.

1.3 Overview

This thesis is organised as follows:

Chapter 2 shows a review and a definition of linear transformations that can be employed for the analysis, parameterisation or compression of biomedical data. We are particularly considering the wavelet transform, the wavelet packet transformation and the Gabor expansion under the aspect of data defined on a finite interval. For this, we introduce a matrix notation for each transformation method. Also, appropriate signal extension methods are described for data on finite intervals.

Chapter 3 deals with the methods we studied and developed for the feature selection. Firstly, a simple energy reduction approach is stated. Then, statistical tests are explained that can be used to increase the significance when only few data points are available. These methods select certain TF coefficients and the separability performance of

each selected coefficient can be evaluated by a receiver operating characteristic (ROC) analysis. The ROC analysis is used to develop a SNR-like criterion, that selects and combines significant TF coefficients to a coefficient set for which a separability can be stated. Also, the found coefficient set can be again evaluated by a ROC analysis.

Chapter 4 explains SVM starting with an introduction to learning theory, followed by the SVM theory. The third section shows how we can use SVM for detection of biomedical signals by introducing a connection to a diagnostic test. Also, multi-class SVM classifiers are stated with the introduction of a neutral class. Moreover, it is shown that the non-linear decision boundary found by the SVM can also be evaluated by a ROC analysis.

Chapter 5 introduces the application of some of the introduced signal processing tools to data from subjects that suffer from panic disorder. The chapter starts with a literature review, comprising analysis and studies that have been conducted on panic disorder so far. The feature selection is shown for statistical tests based on TF transformed data. This approach is confirmed by the use of SVM where better separability results are obtained for the parameterised data than for the unparameterised data.

Chapter 6 comprises the development of a differential diagnosis method for determining cochlear hearing loss based on TF transformed TEOAE data. Firstly, a literature review is given. After this, frequency specific cochlear hearing loss is determined by analysing TEOAE data. By the method, a set of distinctive coefficients are obtained which generalises and enhances the method presented in [13]. Then, SVM are applied for the classification which is evaluated by ROC analysis. The chapter concludes with a comparison of the results for each transformation method and a summary.

Chapter 7 concludes our study and discusses future work.

Chapter 2

Feature Extraction: Time-Frequency Transforms

This chapter reviews and defines linear transformations that can be employed for the analysis, parameterisation or compression of biomedical data. We are particularly considering the wavelet transform, the wavelet packet transformation and the Gabor expansion under the aspect of data defined on a finite interval. For this, Section 2.1 introduces linear discrete transformations and their properties. Section 2.2 describes signal extension methods for data defined on finite intervals only. In Section 2.3 the wavelet transformation is presented. Section 2.4 describes the wavelet packet transformation as a generalisation of the wavelet transformation. The last transformation introduced in Section 2.5 is the Gabor frames transformation. The chapter concludes with a brief summary and discussion in Section 2.6.

2.1 Introduction

2.1.1 Linear Discrete Transformation

A linear discrete transformation can be denoted as

$$\mathbf{y} = \mathbf{H}\mathbf{x}, \tag{2.1}$$

with a coefficient vector $\mathbf{x} = \left[x[0] \;\; x[1] \;\; \ldots \;\; x[N-1]\right]^T$, a transform matrix $\mathbf{H} \in \mathbb{C}^{K \times N}$ and a transform vector $\mathbf{y} = \left[y[0] \;\; y[1] \;\; \ldots \;\; y[K-1]\right]^T$. The vector $\mathbf{x}$ defines the discrete data to be analysed on a finite interval $0 < n < N$. The vector $\mathbf{y}$ contains the discrete transformed coefficients on a finite interval $0 < k < K$. The basic aim of the transformation is to transfer the vector $\mathbf{x}$ in the time domain into a representation that allows us to see its main characteristics in the transform domain.

The matrix $\mathbf{H}$ shall possess the property of being dense, meaning that for every vector $\mathbf{x}$ there exists one unique vector $\mathbf{y}$ in the transform domain. For this, a condition arises for the matrix $\mathbf{H}$ [25]: Its null-space has to be the trivial case $\mathbf{0}$ only, $\text{null}\{\mathbf{H}\} = \mathbf{0}$. This means that $\mathbf{y} \neq \mathbf{0} \; \forall \; \mathbf{x} \neq \mathbf{0}$. If this condition is not met, the transformation is referred to as being not dense.

2.1.2 Inverse Transformation

For the linear discrete transformation to be invertible, the transformation matrix $\mathbf{H}$ must be dense. If $\mathbf{H}$ is dense, an inverse transformation exists. We can distinguish three cases [25]:

1. For $K = N$ and $\text{rank}\{\mathbf{H}\} = N$, the inverse transform matrix for $\mathbf{H}$ equals $\mathbf{H}^{-1}$.
2. For $K > N$ and $\text{rank}\{\mathbf{H}\} = N$, the inverse transform matrix for $\mathbf{H}$ equals $(\mathbf{H}^H \cdot \mathbf{H})^{-1}\mathbf{H}^H$ and is called left pseudo-inverse $\mathbf{H}^\dagger$. For this case, the columns of $\mathbf{H}$ are linearly independent.
3. For $K < N$ the transformation is not dense and hence not invertible.

In the following, we briefly describe the transformations on the "edges" of Figure 1.2. For our studies, these transforms are mainly used for confirmation purposes of our approaches.

2.1.3 Discrete Fourier Transform

A well-known classical linear discrete transformation is the Discrete Fourier Transformation (DFT, [26]). The DFT transformation matrix is given by

$$\mathbf{H}_{DFT} = \frac{1}{N}\begin{bmatrix} a^{0\cdot 0} & a^{0\cdot 1} & \dots & a^{0\cdot(N-1)} \\ a^{1\cdot 0} & a^{1\cdot 1} & \dots & a^{1\cdot(N-1)} \\ \cdot & \cdot & & \cdot \\ \cdot & \cdot & & \cdot \\ \cdot & \cdot & & \cdot \\ a^{(N-1)\cdot 0} & a^{(N-1)\cdot 1} & \dots & a^{(N-1)\cdot(N-1)} \end{bmatrix} \tag{2.2}$$

with $a = e^{-jc\frac{2\pi}{N}}$. The DFT according to $\mathbf{y} = \mathbf{H}_{DFT}\mathbf{x}$ represents the discrete approximation of the continuous Fourier-Integral $Y(\omega) = \int x(t)e^{-jc\omega t}dt$.

The advantages of the DFT are the fast implementation via Fast Fourier Transform (FFT, [27]) and that it gives access to frequency domain information. A disadvantage is that there is no resolution with respect to the time domain. Moreover, we refer to the DFT as a fixed transform, meaning the basis functions can not be adapted to the data that ought be be analysed. We continue with the other "extreme" case according to Figure 1.2.

2.1.4 Karhunen-Loeve Transform

The Karhunen-Loeve transform (KLT, [28]) is based on singular value decomposition (SVD) and hence we firstly introduce the SVD before explaining the KLT. To visualise structures in high-dimensional data, the SVD reduces the dimension and identifies subspaces that capture most of the features in the data and displays an ability to detect relative weak signatures along with achieving a reduction of noise [28]. SVD is also conceptually similar to principal component analysis [29], where a covariance matrix is used as input for the algorithm that conducts the SVD. The explanation of the algorithm as well as a detailed mathematical description is beyond our scope and can be found in [28]. We

introduce the SVD as a noise reduction method that is based on a lemma of linear algebra [25] saying that any matrix $\mathbf{D}$ with the dimension $K \times N$ can be decomposed the following way:

$$\mathbf{D} = \mathbf{U} \cdot \mathbf{S} \cdot \mathbf{V}^T, \tag{2.3}$$

with $\mathbf{U} \in \mathbb{C}^{K \times K}$, $\mathbf{S} \in \mathbb{C}^{K \times N}$ and $\mathbf{V}^T \in \mathbb{C}^{N \times N}$.

In more detail, for a square data matrix with $K = N$ the SVD corresponds to:

$$\mathbf{D} = [\mathbf{u}_1 \ ... \ \mathbf{u}_K] \begin{bmatrix} \sigma_1 & & \\ & \ddots & \\ & & \sigma_K \end{bmatrix} \begin{bmatrix} \mathbf{v}_1^T \\ \vdots \\ \mathbf{v}_N^T \end{bmatrix}, \tag{2.4}$$

with the entries in the diagonal of $\mathbf{S}$, $\sigma_1 \ ... \ \sigma_K$, called the singular values of $\mathbf{D}$. If $\mathbf{D}$ has full rank, $\text{rank}\{\mathbf{D}\} = K$, the singular values are unequal to zero. If $\mathbf{D}$ is rank-deficient, $\text{rank}\{\mathbf{D}\} = k_d$, $k_d < K$, the singular values $\sigma_K ... \ \sigma_{K-k_d}$ equal zero, leading to the following equation:

$$\mathbf{D} = [\mathbf{u}_1 \ ... \ \mathbf{u}_{k_d} \mid \mathbf{u}_{k_d+1} \ ... \ \mathbf{u}_K] \left[\begin{array}{ccc|c} \sigma_1 & & & \\ & \ddots & & 0 \\ & & \sigma_{k_d} & \\ \hline & 0 & & 0 \end{array}\right] \left[\begin{array}{c} \mathbf{v}_1^T \\ \vdots \\ \mathbf{v}_{k_d}^T \\ \hline \mathbf{v}_{k_d+1}^T \\ \vdots \\ \mathbf{v}_N^T \end{array}\right]. \tag{2.5}$$

Also, a rectangular matrix $\mathbf{D}$ with $K > N$ and $\text{rank}\{\mathbf{D}\} = N$ has a SVD of:

$$\mathbf{D} = [\mathbf{u}_1 \ ... \ \mathbf{u}_K] \left[\begin{array}{ccc} \sigma_1 & & \\ & \ddots & \\ & & \sigma_N \\ \hline & 0 & \end{array}\right] \begin{bmatrix} \mathbf{v}_1^T \\ \vdots \\ \mathbf{v}_N^T \end{bmatrix}. \tag{2.6}$$

The explanation for our application is as follows: a original data matrix $\mathbf{D}$ of K samples by N variables is decomposed to a matrix of new orthogonal vectors contained by $\mathbf{V}^T$, representing linear combinations of the original variables, a square matrix $\mathbf{S}$ whose diagonal contains the singular values, and an orthonormal matrix of scores $\mathbf{U}$ for observations in the new orientation.

The product $\mathbf{U} \cdot \mathbf{S}$ with dimension $K \times N$ can be interpreted as the coordinates of each column of $\mathbf{D}$ as a point in a transformed space spanned by the rows of $\mathbf{V}^T$. The transformed space has the property that the maximal possible variation occurs along the axis corresponding to the first column of $\mathbf{U}$, the maximal remaining variation along the axis corresponding to the second column of $\mathbf{U}$ and so on. In most cases, the majority of the noise in a dataset may be eliminated by truncating (2.4) similarly to (2.5) to a k_t-dimensional space $k_t < \text{rank}\{\mathbf{D}\}$ while still providing a close approximation the original data matrix $\mathbf{D}$.

The real biomedical data for our studies usually has a rectangular data matrix as there are more samples than variables. Also, as our data is severely corrupted by noise which is the case for most biomedical data, the data matrix can be assumed to have full rank, similarly to (2.6).

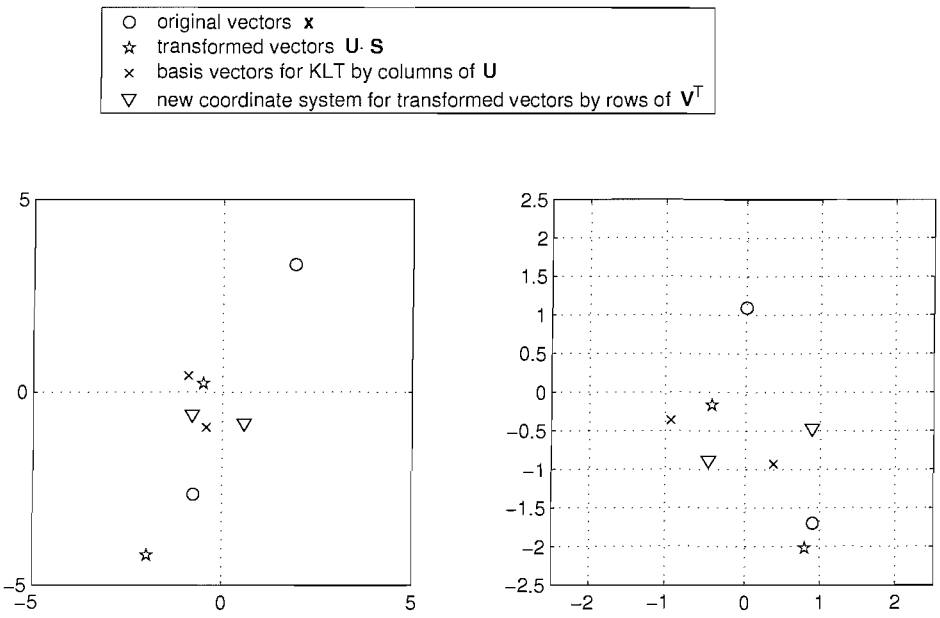

Figure 2.1: Sample SVD decomposition.

To underline the above statements an example illustrated by Figure 2.1 is given. The sample vector $\mathbf{s} = \begin{bmatrix} 2 \\ 3 \end{bmatrix}$ representing the signal whose features are to be identified is changed and corrupted by noise the following way: $\mathbf{x}_1 = -\frac{3}{4} \cdot \mathbf{s} + \begin{bmatrix} 0.75 \\ -0.4 \end{bmatrix}$ and $\mathbf{x}_2 = \mathbf{s} + \begin{bmatrix} -0.1 \\ 0.3 \end{bmatrix}$ which leads to a data matrix $\mathbf{D} = [\mathbf{x}_1 \ \mathbf{x}_2] = \begin{bmatrix} -0.75 & 1.9 \\ -2.65 & 3.3 \end{bmatrix}$ for the left case of Figure 2.1; to illustrate a second example on the right of the figure a signal vector $\mathbf{s} = \begin{bmatrix} 1 \\ -2 \end{bmatrix}$ is changed the same way which leads to $\mathbf{D} = \begin{bmatrix} 0 & 0.9 \\ 1.1 & -1.7 \end{bmatrix}$.

For illustrative purposes the SVD decomposition for the two examples is shown graphically in Figure 2.1. A detailed observation of the figure yields that the plotted columns of $\mathbf{U}$ are orthogonal as well as the illustrated rows of $\mathbf{V}^T$. Furthermore, the most important property of the SVD can be seen: When looking at the the transformed vectors obtained by the product $\mathbf{U} \cdot \mathbf{S}$ for both examples, a relatively good approximation of the respective signal vectors is illustrated by one vector which corresponds to the first column of the product matrix and the noise is mainly contained in the second vector which corresponds to the second column of the product matrix. However, the transformed vectors $\mathbf{U} \cdot \mathbf{S}$ are based on a new coordinate system described by $\mathbf{V}^T$. Also, for the example in Figure 2.1 on the left, the approximation of the signal vector is reflected showing an opposite sign.

To illustrate the property of the SVD for $K > N$, another dimension can be added to the signal vector. $\mathbf{s} = \begin{bmatrix} 2 \\ 3 \\ 3 \end{bmatrix}$, the changes made to the signal vector are: $\mathbf{x}_1 = -\frac{3}{4} \cdot \mathbf{s} + \begin{bmatrix} 0.75 \\ -0.4 \\ 0.85 \end{bmatrix}$ and $\mathbf{x}_2 = \mathbf{s} + \begin{bmatrix} -0.1 \\ 0.3 \\ -1.5 \end{bmatrix}$ which leads to a data matrix $\mathbf{D} = \begin{bmatrix} -0.75 & 1.9 \\ -2.65 & 3.3 \\ -1.4 & 1.5 \end{bmatrix}$. The product matrix for this case

equals: $\mathbf{U} \cdot \mathbf{S} = \begin{bmatrix} -1.97 & 0.54 \\ -4.23 & -0.14 \\ -2.04 & -0.22 \end{bmatrix}$. Again, the first column of the product matrix yields a reasonable approximation of the original signal vector $\mathbf{s} = \begin{bmatrix} 2 \\ 3 \\ 3 \end{bmatrix}$ with an opposite sign in the coordinate system defined by $\mathbf{V}^T$, whereas the second column represents mostly noise. Now, representing the data matrix by only the first column of the product matrix reduces the dimension from $N = 2$ to $k_t = 1$ for this example. This shows the property of the SVD to reduce the dimensions of high-dimensional data.

Now, we can proceed to show the relation of the SVD to the KLT aiming at defining a transformation matrix $\mathbf{H}$ according to (2.1). Simply, based on the above explanations $\mathbf{H}_{KLT}$ is defined as:

$$\mathbf{H}_{KLT} = \mathbf{U}^T, \tag{2.7}$$

meaning the KLT transformation matrix uses the columns of $\mathbf{U}$ as basis vectors. For our examples, we get:

$$\text{For } \mathbf{D} = \begin{bmatrix} -0.75 & 1.9 \\ -2.65 & 3.3 \end{bmatrix}: \quad \mathbf{y}_1 = \mathbf{H}_{KLT} \cdot \mathbf{x}_1 = \mathbf{U}^T \cdot \mathbf{x}_1 = \begin{bmatrix} -0.42 & -0.91 \\ -0.91 & 0.42 \end{bmatrix} \cdot \mathbf{x}_1 = \begin{bmatrix} 2.72 \\ -0.45 \end{bmatrix},$$

$$\mathbf{y}_2 = \mathbf{H}_{KLT} \cdot \mathbf{x}_2 = \mathbf{U}^T \cdot \mathbf{x}_2 = \begin{bmatrix} -0.42 & -0.91 \\ -0.91 & 0.42 \end{bmatrix} \cdot \mathbf{x}_2 = \begin{bmatrix} -3.79 \\ -0.32 \end{bmatrix},$$

$$\text{for } \mathbf{D} = \begin{bmatrix} 0 & 0.9 \\ 1.1 & -1.7 \end{bmatrix}: \quad \mathbf{y}_1 = \mathbf{H}_{KLT} \cdot \mathbf{x}_1 = \mathbf{U}^T \cdot \mathbf{x}_1 = \begin{bmatrix} 0.37 & -0.93 \\ -0.93 & -0.37 \end{bmatrix} \cdot \mathbf{x}_1 = \begin{bmatrix} -1.02 \\ -0.40 \end{bmatrix},$$

$$\mathbf{y}_2 = \mathbf{H}_{KLT} \cdot \mathbf{x}_2 = \mathbf{U}^T \cdot \mathbf{x}_2 = \begin{bmatrix} 0.37 & -0.93 \\ -0.93 & -0.37 \end{bmatrix} \cdot \mathbf{x}_2 = \begin{bmatrix} 1.91 \\ -0.22 \end{bmatrix},$$

$$\text{for } \mathbf{D} = \begin{bmatrix} -0.75 & 1.9 \\ -2.65 & 3.3 \\ -1.4 & 1.5 \end{bmatrix}: \quad \mathbf{y}_1 = \mathbf{H}_{KLT} \cdot \mathbf{x}_1 = \mathbf{U}^T \cdot \mathbf{x}_1 = \begin{bmatrix} -0.39 & -0.83 & -0.40 \\ 0.90 & -0.24 & -0.37 \\ 0.21 & -0.50 & 0.83 \end{bmatrix} \cdot \mathbf{x}_1 = \begin{bmatrix} 3.05 \\ 0.48 \\ 0 \end{bmatrix},$$

$$\mathbf{y}_2 = \mathbf{H}_{KLT} \cdot \mathbf{x}_2 = \mathbf{U}^T \cdot \mathbf{x}_2 = \begin{bmatrix} -0.39 & -0.83 & -0.40 \\ 0.90 & -0.24 & -0.37 \\ 0.21 & -0.50 & 0.83 \end{bmatrix} \cdot \mathbf{x}_2 = \begin{bmatrix} -4.08 \\ 0.36 \\ 0 \end{bmatrix}.$$

It can be observed that the KLT leads to a parameterisation of the data where the specially adapted first basis functions contain most of the signal information, resulting in transformed vectors $\mathbf{y}$ with a very large component at the beginning, decreasing rapidly. This also shows that the KLT is fully adapted to the data. A major drawback of this transform is the generalisation. When control data is used to justify the selected features by the KLT, results can be very poor.

To achieve better generalisation, other transforms need to be employed. Examples for such transforms are the discrete wavelet transformation (DWT, [30]), the wavelet packets (WP, [31]) and Gabor frames (GF, [32]). Here, the transforms are based on short and temporally translated kernel functions that describe the signal information in the data matrix $\mathbf{D}$. They can also be seen as a compromise

between the DFT and the KLT as they are not fully adapted to the data but can be modified based on certain data properties by changing and adapting the kernel functions. In this report the implementation of these transforms based on filter banks is described. We wish to use the information obtained by the kernel functions and the resulting vicinity to filtering also at the end of the finite data vector $\mathbf{x}$. Therefore, it is necessary to suitably extend $\mathbf{x}$.

2.2 Signal Extension

This section deals with the boundary problem that is caused by data defined on finite intervals and shows an appropriate signal extension to solve the problem.

For the linear discrete transformation $\mathbf{y} = \mathbf{H}\mathbf{x}$, we aim to be able to choose having the same or a greater vector length for $\mathbf{y}$ as for $\mathbf{x}$. As the case of having the same length for $\mathbf{y}$ and $\mathbf{x}$ is more restrictive, we focus on this. Our transformation methods are based on finite-length filter banks, and the convolution of a discrete signal of length N with a discrete filter of length N_F yields a signal with length $N + N_F - 1$. In order to reduce this convolution result to N, we need to extend the signal.

Three ways to accomplish such a signal extension are shown in [26]:

- Extension with zeros (zero padding),
- Periodic extension (wraparound),
- Extension by reflection (symmetric extension).

The DFT in 2.1.4 includes a periodic extension. As extension with zeros loses data and periodic extension suffers from blurring features hidden close to the interval margins of the data, we limit ourselves to apply only the extension by reflection to the linear discrete transformations.

For this symmetric extension we separate two cases [26]:
The impulse response of the filter is symmetric and of odd length or
the impulse response of the filter is symmetric and of even length.

The first case is illustrated in Figure 2.2. The figure shows that the convolution of a symmetrically extended discrete signal $\tilde{x}[n]$ (to theoretically $\pm\infty$) with a discrete symmetric filter $h[n]$ of odd length yields a symmetric signal $\tilde{y}[n]$ with the same period as $\tilde{x}[n]$. We just cut out one period of it to find $y[n]$ that has the same length as $x[n]$. The points of symmetry of $\tilde{x}[n]$ are $n = 4, 7, 10, 13$. A requirement for this extension is that the filter $h[n]$ is symmetric. This leads to some constraints as for example some popular wavelets on which the DWT is based are not symmetric and therefore, cannot be extended symmetrically.

The second case for symmetric extension appears when the impulse response of the filter is symmetric and even. This case is shown in Figure 2.3. Here, the convolution of a symmetrically extended discrete signal $\tilde{x}[n]$ (to theoretically $\pm\infty$) with a discrete symmetric filter $h[n]$ of even length yields a symmetric signal $\tilde{y}[n]$ whereby the period of $\tilde{y}[n]$ equals the period of $\tilde{x}[n]$ plus one. Hence, with this symmetric extension the signal $y[n]$ which is obtained by filtering has one element more than $x[n]$. The points of symmetry of $\tilde{x}[n]$ are $n = 4.5, 8.5, 12.5$.

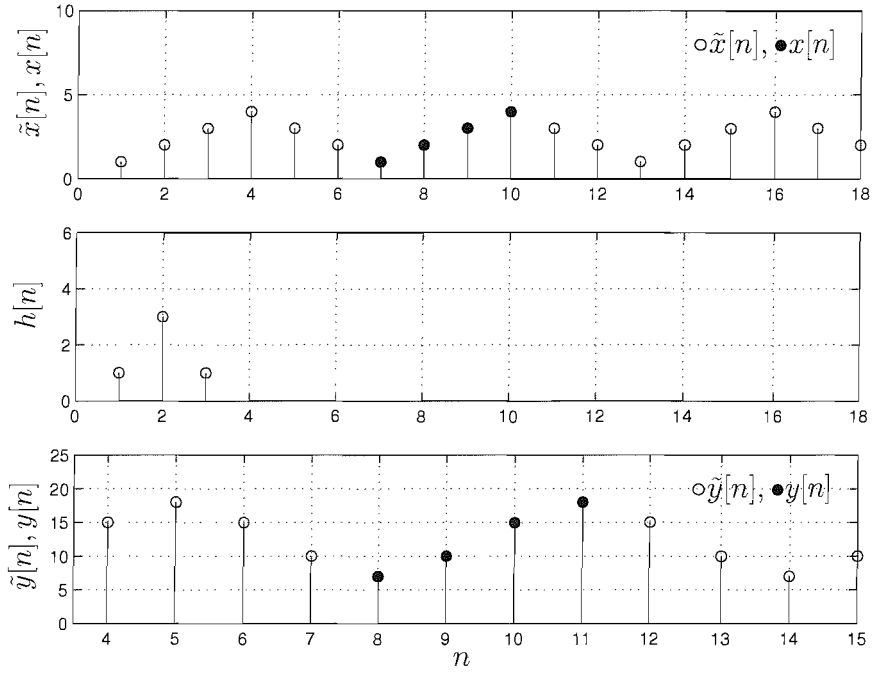

Figure 2.2: Case 1: (top) extension of $x[n]$ yielding $\tilde{x}[n]$, (middle) odd length filter impulse response $h[n]$ and (bottom) convolution result, from which $y[n]$ can be extracted.

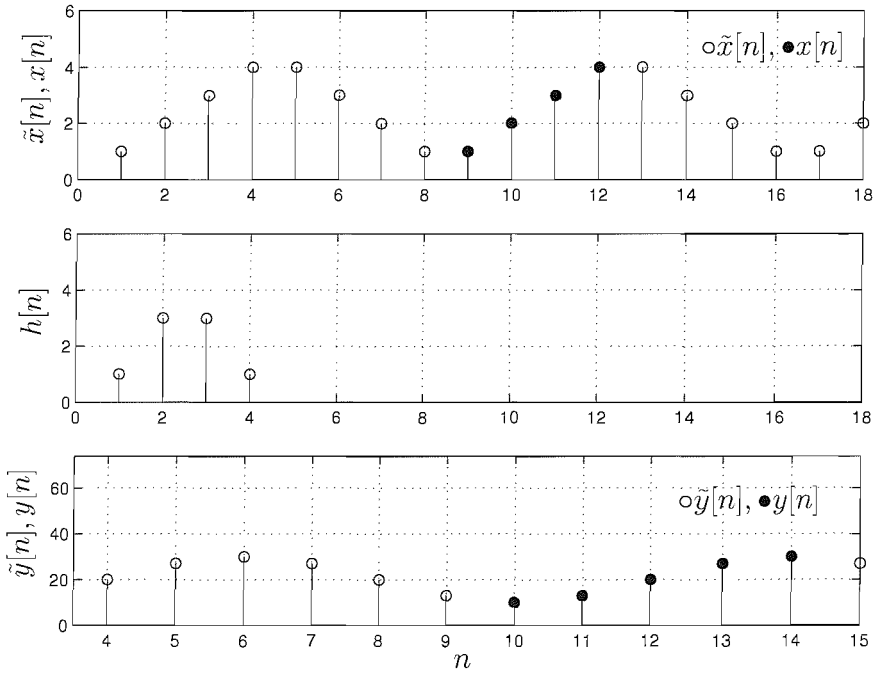

Figure 2.3: Case 2: (top) extension of $x[n]$ yielding $\tilde{x}[n]$, (middle) even length filter impulse response $h[n]$ and (bottom) convolution result, from which $y[n]$ can be extracted.

For the transformations that will be described in the next sections, we aim to incorporate the signal extension in the transformation matrix **H**. The examples for the DWT and WP do also have a symmetric extension incorporated which cannot be observed as these examples are special cases. In Subsection 2.5.3 it is illustrated in detail how the signal extension is incorporated in the transformation matrix **H** which we applied exactly the same way to the DWT and WP transforms.

2.3 Wavelets

This section reviews the Wavelet Transformation for both the continuous and the discrete case [33], [34]. Then, the matrix representation of the discrete wavelet transformation is introduced.

2.3.1 Continuous Wavelet Transformation

The continuous wavelet transformation (CWT) for a function $x(t)$ is defined as [27]

$$(W_\Psi x)(b,a) = \int x(t)\Psi_{b,a}(t)dt, \tag{2.8}$$

with

$$\Psi_{b,a}(t) = \frac{1}{\sqrt{|a|}}\Psi\left(\frac{t-b}{a}\right) \tag{2.9}$$

for $a, b \in \mathbb{R}$, $a \neq 0$. The function $\Psi(t)$ is called mother wavelet, which can be scaled and translated by the parameters a and b respectively. The mother wavelet is a prototype from which all wavelets used in the transformation are derived by scaling or translation. The term "wavelet" means local wave and refers to the finite support of $\Psi(t)$. The wavelets are the basis functions for the CWT. This is in contrast to the Fourier Transform (see (2.2)) where the basis consists of sines and cosines. These are perfectly localised in frequency space but do not decay as a function of time; wavelets on the other hand decay to zero as $t \to \pm\infty$ and show good localisation properties in the frequency domain. Therefore, wavelets are better suited to represent functions that are localised both in time and frequency.

The CWT according to (2.8) is redundant. Furthermore, there exists an inverse transform that is not clear [35]. To obtain a transform which can be implemented on a computing device, a discretisation of (2.8) is necessary.

2.3.2 Discrete Wavelet Transformation (DWT)

The discretisation of the parameters a, b in the transform domain leads to the discrete wavelet transformation (DWT). For a still continuous function $x(t)$ the dyadic segmentation of the basis system is of interest. Hence, the dyadic discrete wavelet basis system is

$$\Psi_{j,k}[n] = 2^{-j/2}\Psi[2^{-j}n - k] \qquad j, k \in \mathbb{Z}. \tag{2.10}$$

We see that for a dyadic lattice $a = 2^j, b = 2^j k$. As we analyse discrete data, we need to define the DWT for a discretisation in the time domain, leading to a approximation of the continuous case. The

DWT for a discrete function $x[n]$ results in a modification to (2.8) as

$$(W_\Psi x)[j,k] = 2^{-j/2} \sum_{n \in \mathbb{Z}} x[n] \Psi[2^{-j}n - k]. \tag{2.11}$$

An efficient calculation of the DWT coefficients in the case of discrete-time data can be achieved with a multi-resolution algorithm (MRA, [30]). This MRA is performed by filtering the function to be analysed with an octave filter bank as shown in Figure 2.4. The high-pass filter $h_H[n]$ forms a quadrature mirror filter (QMF) pair [36] with the low-pass $h_L[n]$ of the filter bank. The input sequence to the octave filter bank, $x[n]$, is the function to be analysed. Through successive low- and high-pass filtering of the samples in the lower frequency band, $y_{i_j}[k]$, and decimation of the resulting signals by a factor of 2 (denoted as $\downarrow 2$), subband samples $y_{d_j}[k]$ are obtained, which, except of the lowest frequency band, represent the DWT coefficients and contain the detail information of $x[n]$. The coefficients $y_{i_j}[k]$ are intermediate values and correspond to a dual basis function of the wavelet, a so called scaling function [30]. The filters $h_L[n]$ and $h_H[n]$ are sampled version of the underlying scaling function and wavelet, and therefore determine which DWT — amongst a large variety of possible wavelet functions (see e.g. [30, 31, 37]) — is being implemented.

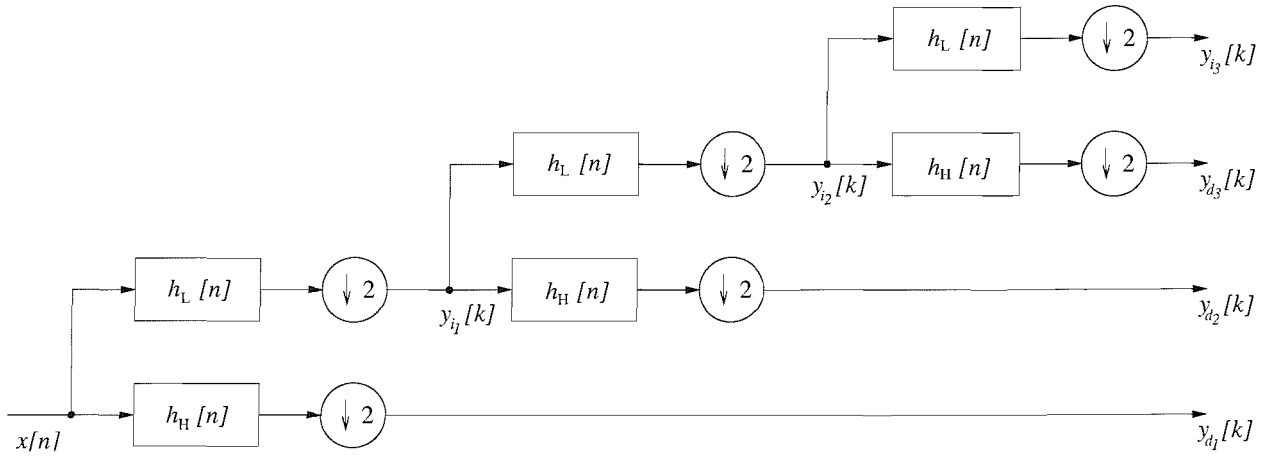

Figure 2.4: Octave filter bank to compute an MRA of depth $J = 3$; deeper decompositions are achieved by further splitting $y_{i_3}[k]$.

According to Figure 2.4, the scaling function follows the convolution equation:

$$y_{i_{j+1}}[k] = 2^{1/2} \sum_n y_{i_j}[n] h_L[2k - n]. \tag{2.12}$$

Analogous to that, the wavelet equation results in

$$y_{d_{j+1}}[k] = 2^{1/2} \sum_n y_{i_j}[n] h_H[2k - n] \tag{2.13}$$

where the coefficients of the high-pass filter $h_H[n]$ are called wavelet coefficients.

The inverse transformation is obtained by

$$x[n] = \sum_j \sum_k (W_\Psi x)[j,k] \Psi_{j,k}[n]. \tag{2.14}$$

The scaled and translated wavelets have an advantageous property that makes the inverse transformation unique and that, combined with a symmetric extension, leads to the same signal length for the original and transformed signal. They are orthonormal meaning that

$$\langle \Psi_{j,k}[n], \Psi_{r,s}[n] \rangle = \delta_{j,r}\delta_{k,s}, \qquad \forall j,k,r,s \in \mathbb{Z} \tag{2.15}$$

where $\delta_{j,r}$ is the Kronecker symbol which is defined as

$$\delta_{j,r} = \begin{cases} 1 & \text{for } j = r; \\ 0 & \text{for } j \neq r. \end{cases} \tag{2.16}$$

In the following, the meaning of one transformation coefficient is described more closely. According to (2.11), inner products [27] between $x[n]$ and the analysing wavelets $\Psi[2^{-j}n - k]$ are conducted to obtain $(W_\Psi x)[j,k]$. We calculate one coefficient for one specific scale j and translation k as

$$(W_\Psi x)[j,k] =< \Psi_{j,k}[n], x[n] > . \tag{2.17}$$

The inner product can be interpreted as a measure of similarity. For two discrete functions with their energy content normalised to 1, the inner product equals

- 0, if they are orthogonal;
- 1, if they are equal.

It conducts a least squares fit of one discrete function to the other. Figure 2.5 shows a sample least squares fit. Hereby, the vector $\mathbf{x}$ corresponds to the finite discrete function $x[n]$, the vectors $\mathbf{h}_1, \mathbf{h}_0$ represent two analysing wavelets $\Psi_{j,k}[n]$. The underlying basis or coordinate system for $\mathbf{x}$ is $\{\mathbf{e}_1, \mathbf{e}_0\}$.

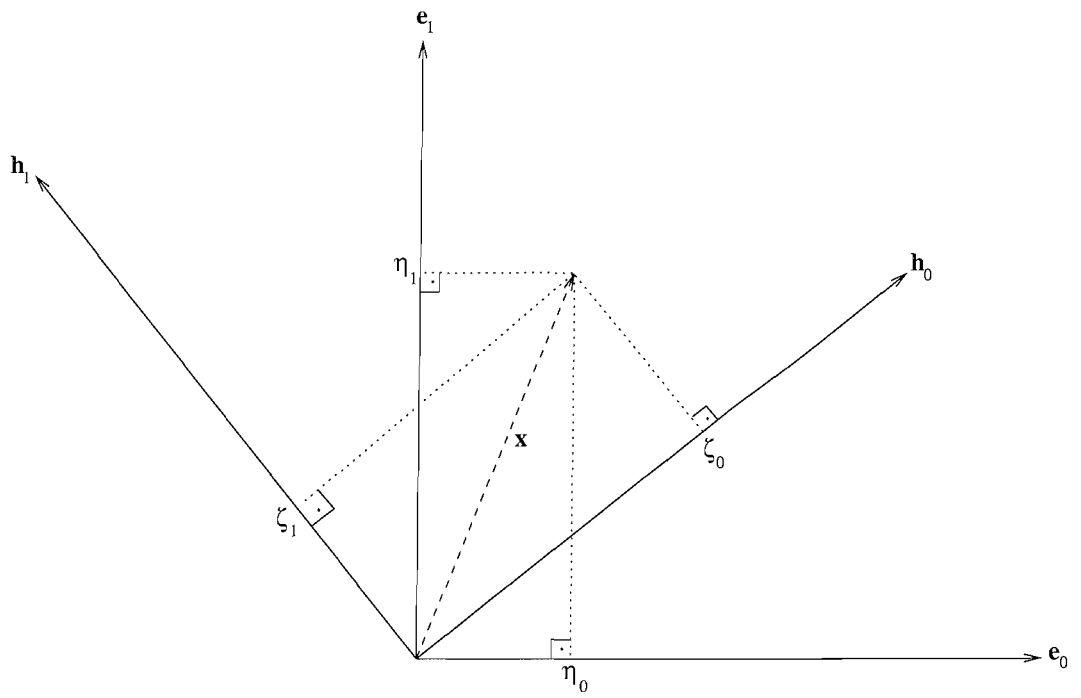

Figure 2.5: Least squares fit of $\mathbf{x}$ onto $\mathbf{h}_1, \mathbf{h}_0$.

The vector $\mathbf{x}$ equals

$$\mathbf{x} = \begin{bmatrix} \eta_1 \\ \eta_0 \end{bmatrix} = \eta_1 \cdot \mathbf{e}_1 + \eta_0 \cdot \mathbf{e}_0 = \zeta_1 \cdot \mathbf{h}_1 + \zeta_0 \cdot \mathbf{h}_0, \tag{2.18}$$

where ζ_1 and ζ_0 are the orthogonal projections of $\mathbf{x}$ onto the new basis $\{\mathbf{h}_1, \mathbf{h}_0\}$. The vector $\mathbf{y}$ based on the new basis $\{\mathbf{h}_1, \mathbf{h}_0\}$ equals

$$\mathbf{y} = \begin{bmatrix} \zeta_1 \\ \zeta_0 \end{bmatrix} = \begin{bmatrix} \mathbf{h}_1^T \cdot \mathbf{x} \\ \mathbf{h}_0^T \cdot \mathbf{x} \end{bmatrix} = \begin{bmatrix} \mathbf{h}_1^T \\ \mathbf{h}_0^T \end{bmatrix} \cdot \mathbf{x} \tag{2.19}$$

where $\mathbf{h}_1$ and $\mathbf{h}_0$ are expressed in terms of the basis $\{\mathbf{e}_1, \mathbf{e}_0\}$.

The vectors $\mathbf{h}_1, \mathbf{h}_0$ represent two basic functions for the transformation. A least squares fit is conducted with the vector $\mathbf{x}$. The vector $\mathbf{y}$ shows how good the basic functions match the vector $\mathbf{x}$. For the DWT, we have N (vector length of $\mathbf{x}$) basic functions which are the wavelets and one scaling function and for each wavelet and the scaling function, a least squares fit with the data vector $\mathbf{x}$ is conducted. Again, $\mathbf{y}$ shows how good the basic functions match the data vector. According to (2.15) the wavelet basic functions are orthonormal. Hence, we can address the DWT as a rotation of the coordinate system of $\mathbf{x}$ with the purpose that $\mathbf{y}$ which is defined by the coordinate system generated by the matrix $\mathbf{H}_{DWT}$ gives us more insight on the characteristics of $\mathbf{x}$.

Let us have a closer look at the dyadic lattice mentioned above. It reveals the advantageous property of the DWT which is that the vector $\mathbf{x}$ is analysed both in time and frequency. Figure 2.6 shows the time-frequency lattice for the DWT. The term frequency and not scale is used because the DWT can be implemented as a filter bank as explained above and the different scales equal different frequency bands [30]. Each rectangle in the figure represents one coefficient of the transform vector $\mathbf{y}$ and shows the zone in the time-frequency plane that is approximately covered by the coefficient.

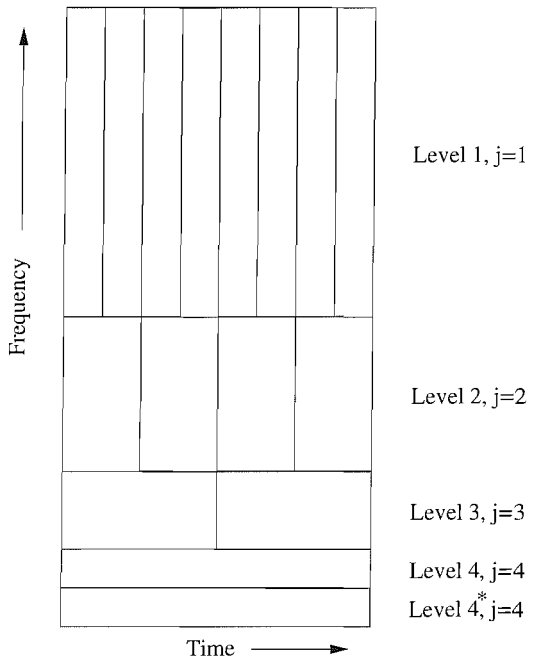

Figure 2.6: Time-frequency tiling for DWT.

The "Level" expressions specify frequency ranges which are created by consecutive filtering with the high-pass and low-pass filters according to (2.12) and (2.13) (see [38]). The Levels 1 to 4 show the

detail information of the signal to be analysed. The Level 4* shows the part of the signal created by the scaling function. High frequencies are resolved poorly in frequency but quite accurately in time. For low frequencies it is the other way around. They are resolved well in frequency but poorly in time. This characteristic is very useful when the signal to be analysed has high frequency components for short durations and low frequency components for long durations which is the case for most medical data.

2.3.3 DWT Transformation Matrix $\mathbf{H}_{DWT}$

We proceed to how we derive the transformation matrix $\mathbf{H}_{DWT}$. Starting from (2.11), one decomposition level j is represented as

$$(W_\Psi x)[j,k] = \sum_{n\in\mathbb{Z}} x[n]h_{o_j}[n-k]$$

with $o = \{d, i\}$ and N equals the length of the discrete function $x[n]$. For $o = d$, the equation for the filter coefficients that reveal the detail information of the analysed signal becomes:

$$h_{d_j}[n] = 2^{-j/2}\Psi[2^{-j}n].$$

For $o = i$ and a maximal decomposition depth of $J = \log_2 N$, the equation that represents the "deepest" level becomes a constant c:

$$h_{i_J}[n] = c.$$

The general structure of the transform vector $\mathbf{y}$ can be derived from Figure 2.4:

$$\mathbf{Y} = \begin{bmatrix} y_{i_J}[k] \\ y_{d_J}[k] \\ \cdot \\ \cdot \\ \cdot \\ y_{d_1}[k] \end{bmatrix}. \tag{2.20}$$

By expressing the discrete functions in vector notation, $y_{o_j}[k] = \mathbf{y}_{o_j}^T$, we arrive at:

$$\mathbf{y} = \begin{bmatrix} \mathbf{y}_{i_J} \\ \mathbf{y}_{d_J} \\ \cdot \\ \cdot \\ \cdot \\ \mathbf{y}_{d_1} \end{bmatrix}. \tag{2.21}$$

Hence, the transformation equation can be written as:

$$\mathbf{y} = \mathbf{H}_{DWT}\mathbf{x} = \begin{bmatrix} \mathbf{H}_{i_J}^{K\times N} \\ \mathbf{H}_{d_J}^{K\times N} \\ \cdot \\ \cdot \\ \cdot \\ \mathbf{H}_{d_1}^{K\times N} \end{bmatrix} \mathbf{x} \tag{2.22}$$

with $K = N/2^j$. It is worth mentioning that for $j = J$, $K = 1$. This means that the matrices $\mathbf{H}_{d_J}$ and $\mathbf{H}_{i_J}$ are row vectors and therefore, the vectors $\mathbf{y}_{d_J}$ and $\mathbf{y}_{i_J}$ contain only one element.

We proceed to the definition of a sample matrix $\mathbf{H}_{d_j}$. For a arbitrary N and $j = 1$, the matrix becomes:

$$\mathbf{H}_{d_1} = \begin{bmatrix} & & & & \vdots & & & & \\ 0 & h_{d_j}[0] & h_{d_j}[1] & \dots & h_{d_j}[n] & 0 & 0 & 0 & \dots \\ 0 & 0 & 0 & h_{d_j}[0] & h_{d_j}[1] & \dots & h_{d_j}[n] & 0 & \dots \\ & & & & \vdots & & & & \end{bmatrix} \tag{2.23}$$

Note that the translation for the successive rows equals 2^j and hence, in our example for $j = 1$, it is 2. This shows how one row in the Matrix $\mathbf{H}_{DWT}$ is defined. The incorporation of the signal extension is discussed later in the section about Gabor frames because of didactic reasons. The listed examples for the DWT and the Wavelet Packets are special cases and therefore can be seen as having a signal extension incorporated that cannot be observed.

To illustrate the above statements the transformation equation with the matrix $\mathbf{H}_{DWT}$ for a Haar wavelet [38] is shown:

$$\begin{bmatrix} \mathbf{y}_{i_3} \\ \mathbf{y}_{d_3} \\ \mathbf{y}_{d_2} \\ \mathbf{y}_{d_1} \end{bmatrix} = \begin{bmatrix} y_1 \\ y_2 \\ y_3 \\ y_4 \\ y_5 \\ y_6 \\ y_7 \\ y_8 \end{bmatrix} = \begin{bmatrix} \frac{1}{\sqrt{8}} & \frac{1}{\sqrt{8}} & \frac{1}{\sqrt{8}} & \frac{1}{\sqrt{8}} & \frac{1}{\sqrt{8}} & \frac{1}{\sqrt{8}} & \frac{1}{\sqrt{8}} & \frac{1}{\sqrt{8}} \\ \frac{1}{\sqrt{8}} & \frac{1}{\sqrt{8}} & \frac{1}{\sqrt{8}} & \frac{1}{\sqrt{8}} & -\frac{1}{\sqrt{8}} & -\frac{1}{\sqrt{8}} & -\frac{1}{\sqrt{8}} & -\frac{1}{\sqrt{8}} \\ \frac{1}{2} & \frac{1}{2} & -\frac{1}{2} & -\frac{1}{2} & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & \frac{1}{2} & \frac{1}{2} & -\frac{1}{2} & -\frac{1}{2} \\ \frac{1}{\sqrt{2}} & -\frac{1}{\sqrt{2}} & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & \frac{1}{\sqrt{2}} & -\frac{1}{\sqrt{2}} & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & \frac{1}{\sqrt{2}} & -\frac{1}{\sqrt{2}} & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & \frac{1}{\sqrt{2}} & -\frac{1}{\sqrt{2}} \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ x_3 \\ x_4 \\ x_5 \\ x_6 \\ x_7 \\ x_8 \end{bmatrix}. \tag{2.24}$$

The signal extension cannot be observed because of the specific properties of the Haar wavelet. The equation shows that the rows contain the coefficients of the wavelets. They are generated by scaling or translation of the preceding row. When considering the DWT as a filter bank, the rows contain the impulse responses of the filters $h_{o_j}[n]$. Therefore, the transformation $\mathbf{y} = \mathbf{H}_{DWT}\mathbf{x}$ can be regarded as a filtering of $\mathbf{x}$ through digital filters $h_{o_j}[n]$ that corresponds to the mathematical operation of convolution with the impulse response of the filters. Hence, the elements of $\mathbf{y}$ contain the components of $\mathbf{x}$ that are approximately fixed by the wavelets $h_{d_j}[n]$ and the constant filter $h_{i_J}[n]$ as shown in Figure 2.7, where the Level 3* contains an approximation of the constant component of the signal.

2.4 Wavelet Packets (WP)

We start this section by explaining how and why we proceed from the DWT to the wavelet packet transform. Then, we give an example of a wavelet packet transformation matrix $\mathbf{H}_{WP}$.

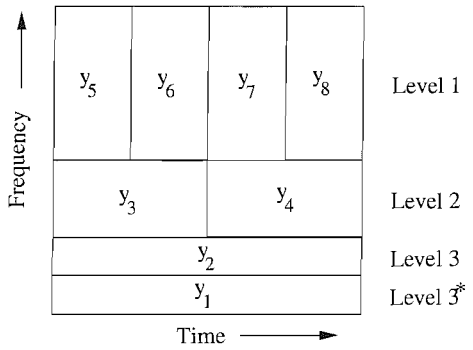

Figure 2.7: Time-frequency segmentation of the vector $\mathbf{y}$ for the DWT.

2.4.1 From Wavelets to Wavelet Packets

The explained dyadic tiling of the time-frequency plane is fixed for the case of the DWT as shown in the previous section. To adapt the transformation to specific data properties, it may be advantageous to further decompose for example level 2 in Figure 2.6 into level 3. This adaptive approach is called Wavelet Packet (WP) decomposition.

As an example, Figure 2.8 depicts a comparison between a wavelet and a sample WP decomposition. The above mentioned properties for wavelets stay the same for WP for example the orthonor-

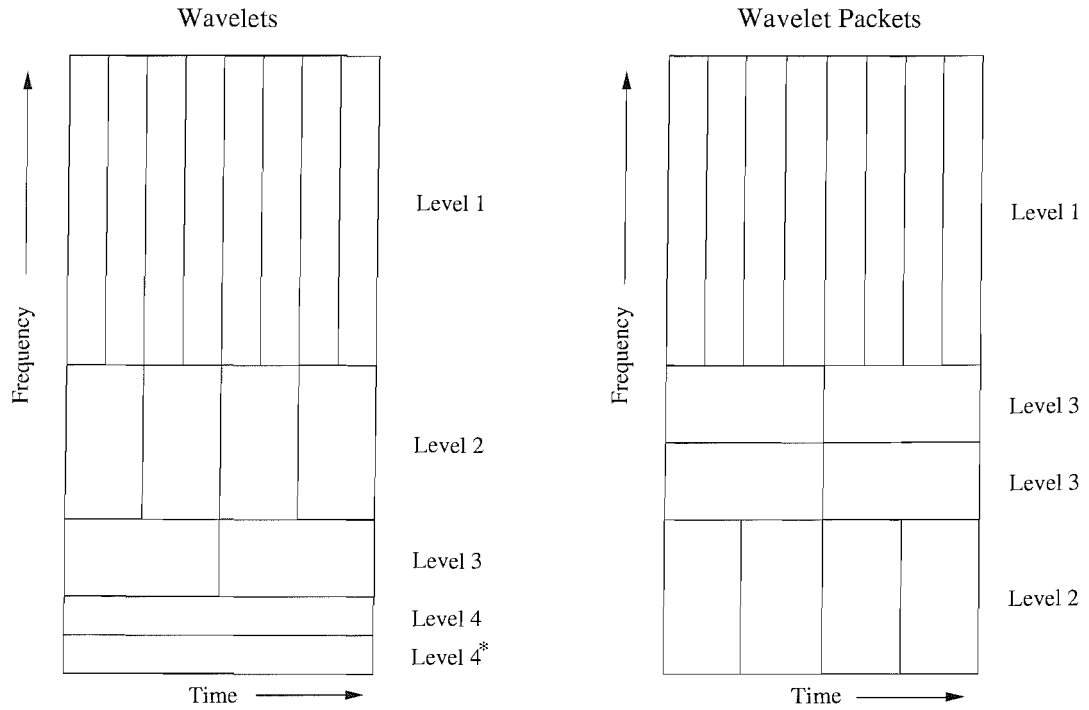

Figure 2.8: Time frequency tiling comparison between a Wavelet and a sample WP decomposition.

mality of the transformation matrix $\mathbf{H}_{WP}$. The difference between the wavelet matrix $\mathbf{H}_{DWT}$ and the wavelet packet matrix $\mathbf{H}_{WP}$ lies in the change of the rows. The rows that contain the level 2 coefficients in $\mathbf{H}_{DWT}$ are replaced by level 3 coefficients in $\mathbf{H}_{WP}$ in our example in Figure 2.8.

2.4.1.1 Entropy Criterion

One could ask the following question at this point: What criterion is used to find the optimal decomposition with WP? The answer is that we try to concentrate the energy of the vector $\mathbf{x}$ in as few coefficients as possible. One suitable measure for such a concentration is given by Shannon's Entropy [38],

$$\epsilon(\mathbf{x}_{norm}) = -\sum_{n=0}^{N-1} \ln({x_{norm}}^2[n]) \cdot ({x_{norm}}^2[n]) \ , \tag{2.25}$$

where ln is the natural logarithm and $\mathbf{x}_{norm}$ is the vector that represents the discrete function $x[n]$ divided by its Euclidean vector norm $||\mathbf{x}||$. ϵ is measure for the concentration of the energy of a vector. The reason why we normalise the vector $\mathbf{x}$ is explained with an example. The entropies for the following unit energy vectors are:

$$\epsilon\left(\frac{1}{2}\begin{bmatrix}1\\1\\1\\1\end{bmatrix}\right) \approx 1.3863, \quad \epsilon\left(\frac{1}{\sqrt{3}}\begin{bmatrix}1\\1\\1\\0\end{bmatrix}\right) \approx 1.0986, \quad \epsilon\left(\frac{1}{\sqrt{2}}\begin{bmatrix}1\\1\\0\\0\end{bmatrix}\right) \approx 0.6931, \quad \epsilon\left(\begin{bmatrix}1\\0\\0\\0\end{bmatrix}\right) = 0.$$

We see that the entropy decreases from a maximum value for a vector where the energy is evenly distributed over all coefficients to zero for a vector where all the energy is concentrated in one coefficient. To illustrate this by a vector with greater length, Figure 2.9 depicts the run of $\epsilon(\mathbf{x})$ over the number of ones contained in $\mathbf{x}$ for a length of $\mathbf{x}$ of 512. If not normalised vectors are used the entropy equation will result in:

$$\epsilon(c \cdot \mathbf{x}) = -\sum_{n=0}^{N-1} \ln(c \cdot x^2[n]) \cdot (c \cdot x^2[n]) = -\sum_{n=0}^{N-1} c \cdot x^2[n] \cdot (\ln(x^2[n]) + \ln(c)) = c \cdot \epsilon(x[n]) + x^2[n] \cdot \epsilon(c),$$

with c being a constant.

Hence, as we want to measure concentration in as few elements as possible, $\mathbf{x}$ needs to be normalised by its energy prior to calculating ϵ. According to [38], Shannon's Entropy follows a ln function for a normalised signal vector $\mathbf{x}$.

Finding the optimal decomposition with WP using the minimisation of Shannon's Entropy works as follows: the entropy of a decomposition level is calculated. Then, the vector is decomposed into the next level and the entropy is calculated again. If the entropy of the higher level is smaller than the entropy of the lower level, the vector is decomposed into the higher level and the above procedure is repeated; if not the decomposition is stopped and the lower level decomposition is adopted. Figure 2.10 shows how the sample decomposition in Figure 2.8 is obtained. The entropy of the vector $\mathbf{x}$ is calculated as ϵ_0. Then, the entropy ϵ_1 of the decomposed $\mathbf{x}$ into level 1 is compared with ϵ_0. As $\epsilon_0 > \epsilon_1$, $\mathbf{x}$ is decomposed into level 1. In the next step, the entropies for the two level 1 decompositions $\epsilon_{1,1}, \epsilon_{1,2}$ are compared with the respective entropies $\epsilon_{2,1}, \epsilon_{2,2}$ of the level 2 decompositions. One level 1 decomposition is further decomposed into level 2 because its entropy is larger than the one of the respective level 2. The decomposition of the other level 1 is stopped because its entropy is smaller than the entropy of the respective level 2. The procedure is repeated for level 3 and we arrive at the decomposition that is illustrated in Figure 2.8. For the decomposition into level 3, the entropy is calculated for the sub-segments only, and not for the entire vector in the example. Hereby, $\epsilon_{3,1}$

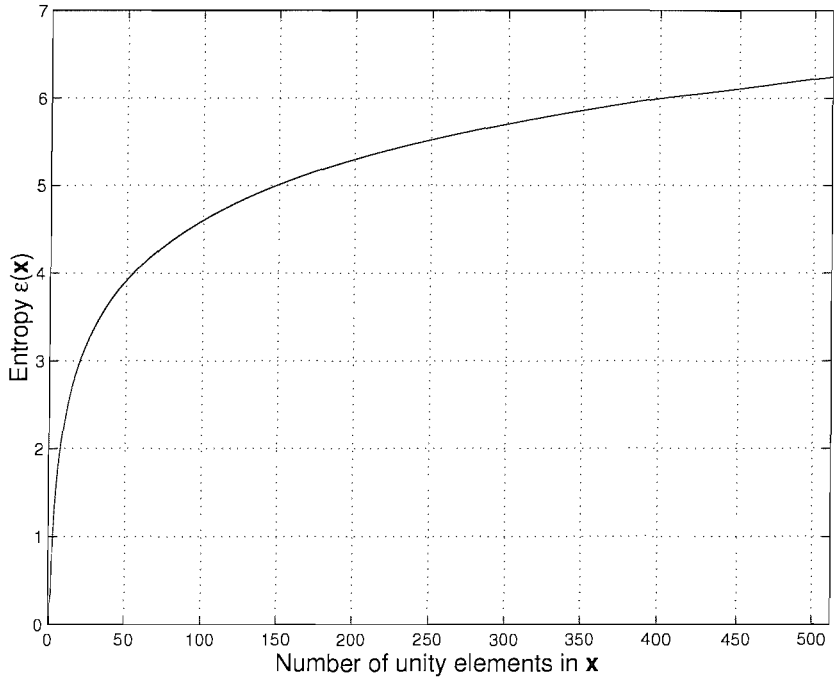

Figure 2.9: Example of the entropy as a function of the sparseness of $\mathbf{x}$ yielding a ln function for a normalised $\mathbf{x}$ in $\epsilon(\mathbf{x})$.

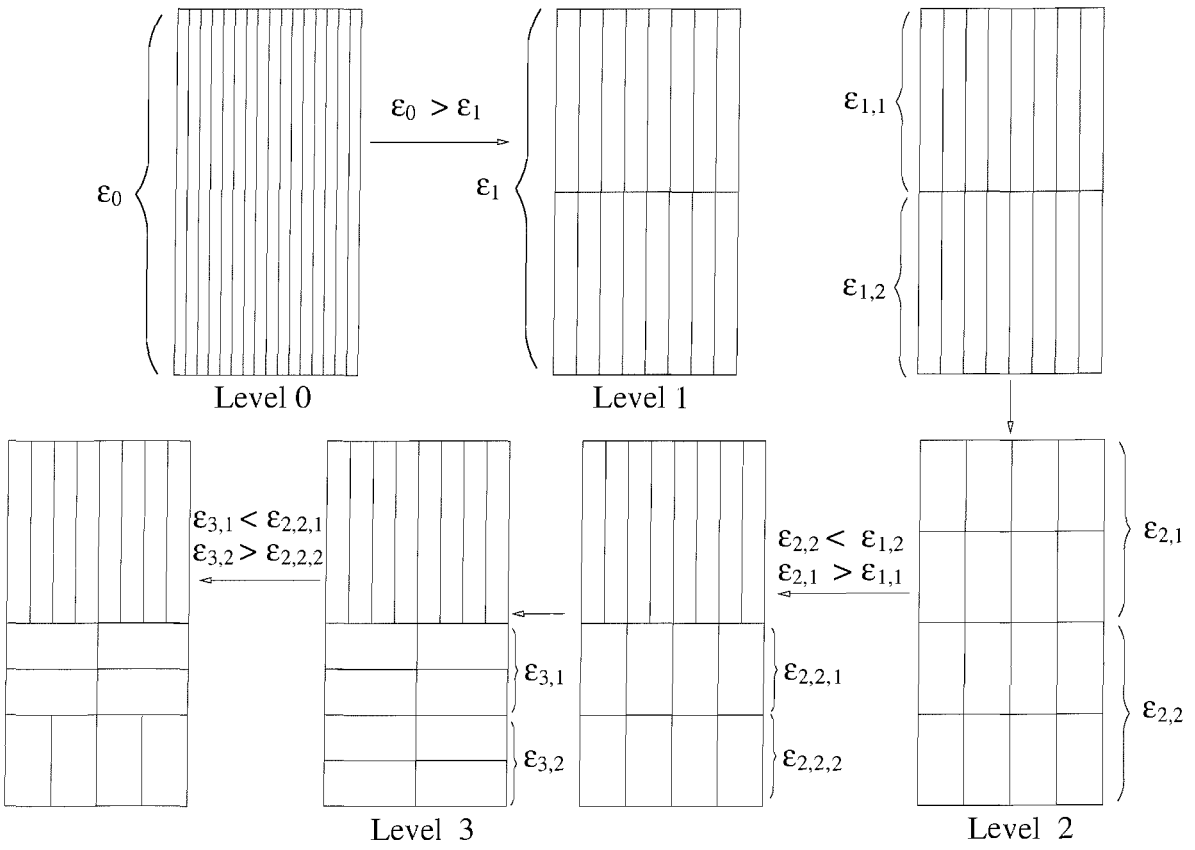

Figure 2.10: WP decomposition of a vector by minimising its entropy.

and $\epsilon_{3,2}$ are compared with $\epsilon_{2,2,1}$ and $\epsilon_{2,2,2}$. The correctness of the procedure is ensured because as Figure 2.9 shows, if only one element of $\mathbf{x}$ is decreased (towards zero), the entropy of the entire vector decreases. The implementation of this procedure is described in the following.

2.4.1.2 Implementation of the Entropy Procedure

To conduct the WP analysis based on the entropy method, we define the data matrix $\mathbf{X}_{\text{data}}^{L\times N}$ where its rows contain the time-domain biomedical data to be analysed. Hence, N defines the length of the finite discrete data and L is the number of data or measurements. The entropy procedure $\mathcal{ENT}$ described above is based on $\mathbf{X}_{\text{data}}$ and yields a decomposition matrix $\mathbf{Z}$ with which the WP coefficients are determined according to

$$\mathcal{WP}_{coeffs},\ \mathbf{Z} = \mathcal{ENT}(\mathbf{X}_{\text{data}}) \tag{2.26}$$

where $\mathbf{Z}$ has the same dimension as $\mathbf{X}_{\text{data}}$. The matrix $\mathbf{Z}$ contains the structure of the WP decomposition for each data vector in its rows according to

$$\mathbf{Z} = \begin{bmatrix} \mathbf{z}_1^T \\ \mathbf{z}_2^T \\ \cdot \\ \cdot \\ \cdot \\ \mathbf{z}_L^T \end{bmatrix}.$$

Figure 2.11 shows a sample row vector $\mathbf{z}_l^T$ of $\mathbf{Z}$ for $N = 16$. It also shows the structure of the DWT vector $\mathbf{z}_{DWT}^T$ for comparison. The vector $\mathbf{z}_l^T$ contains the level number of each coefficient sequentially according to the TF plane.

The row vector $\mathbf{z}_l^T$ in $\mathbf{Z}$ defines the structure of the transformation matrix $\mathbf{H}_{WP}$. Therefore, we have L possible $\mathbf{H}_{WP}$ matrices. Among these we need to find the one that minimises the entropy for all data. Therefore, for each decomposition according to $\mathbf{z}_l^T$ we calculate the entropy $\epsilon_{\mathbf{z}_l^T}$ of $\mathbf{X}_{data}$ and choose the smallest. This procedure is implemented as follows: Start a loop and execute the following for each $\mathbf{z}_l^T$:

- Generate the matrix $\mathbf{W}_{\mathbf{z}_l^T}$ as

$$\mathbf{W}_{\mathbf{z}_l^T}^{N\times L} = \mathbf{H}_{WP_{\mathbf{z}_j^T}}^{N\times N} \cdot (\mathbf{X}_{\text{data}}^{L\times N})^T \tag{2.27}$$

- Determine the entropy for each column of $\mathbf{W}_{\mathbf{z}_l^T}$.
- Calculate the entropy $\epsilon_{\mathbf{z}_l^T}$ by adding up the previously determined entropies in the columns of $\mathbf{W}_{\mathbf{z}_l^T}$ and save $\epsilon_{\mathbf{z}_l^T}$.
- At the end of the loop, choose the smallest $\epsilon_{\mathbf{z}_l^T}$ and denote the corresponding $\mathbf{z}_l^T$ as $\mathbf{z}_{\text{opt}}^T$.

With this procedure we arrive at a WP decomposition, which is specially adapted to the matrix $\mathbf{X}_{\text{data}}$.

$$\mathbf{z}_{DWT}^T = \begin{bmatrix} 4 & 4 & 3 & 3 & 2 & 2 & 2 & 2 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 \end{bmatrix}$$

$$\mathbf{z}_{l\,WP}^T = \begin{bmatrix} 2 & 2 & 2 & 2 & 3 & 3 & 3 & 3 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 \end{bmatrix}$$

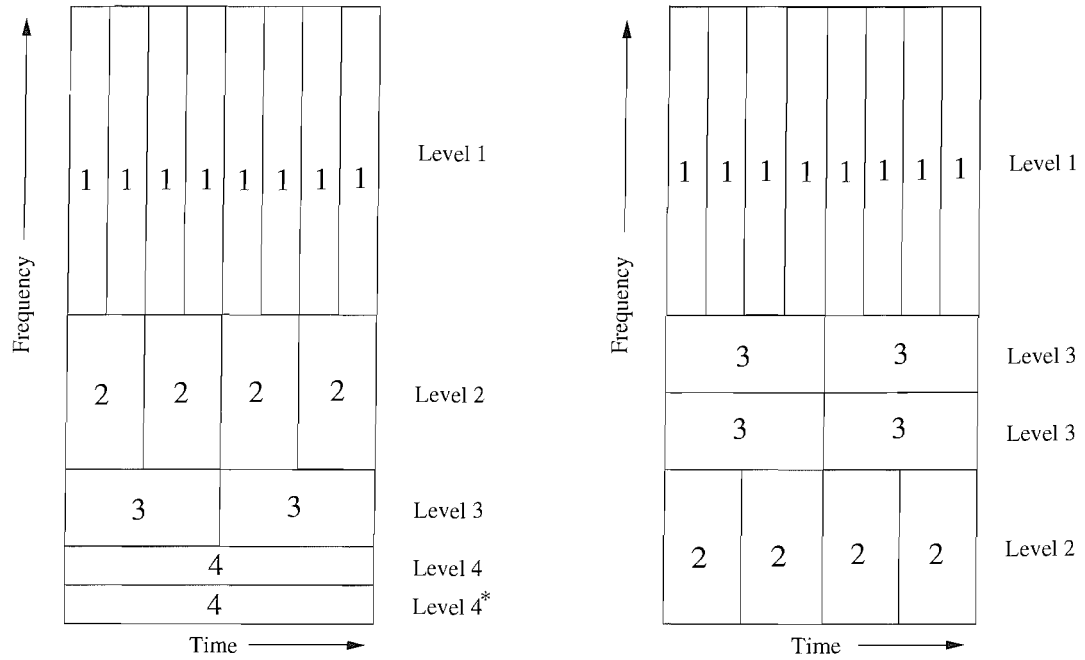

Figure 2.11: Structure comparison and respective TF tiling between (right) a sample WP vector and (left) the DWT vector.

2.4.2 WP Transformation Matrix $\mathbf{H}_{WP}$

In this subsection, an example for a WP matrix $\mathbf{H}_{WP}$ is given.

For instance, the entropy procedure $\mathcal{ENT}$ of a sample data matrix $\mathbf{X}_{\text{data}}$ yields the optimal decomposition vector $\mathbf{z}_{\text{opt}}^T = [2\ 2\ 2\ 2\ 1\ 1\ 1\ 1]$ what defines the WP matrix $\mathbf{H}_{WP}$ and leads to a time-frequency tiling of $\mathbf{y}$ as shown in Figure 2.12 and a corresponding filter bank as shown in Figure 2.13. The

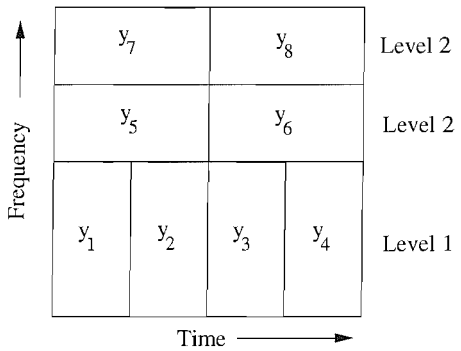

Figure 2.12: Time-frequency segmentation of the vector $\mathbf{y}$ for a sample WP decomposition.

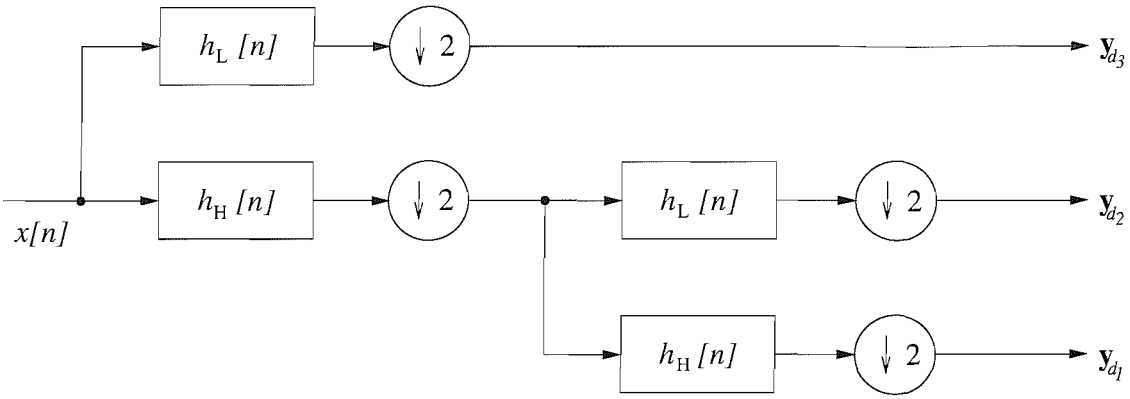

Figure 2.13: Filter bank for the vector $\mathbf{y}$ for a sample WP decomposition.

corresponding transformation equation for a Haar wavelet becomes

$$\begin{bmatrix} \mathbf{y}_{d_3} \\ \mathbf{y}_{d_2} \\ \mathbf{y}_{d_1} \end{bmatrix} = \begin{bmatrix} y_1 \\ y_2 \\ y_3 \\ y_4 \\ y_5 \\ y_6 \\ y_7 \\ y_8 \end{bmatrix} = \begin{bmatrix} \frac{1}{\sqrt{2}} & \frac{1}{\sqrt{2}} & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & \frac{1}{\sqrt{2}} & \frac{1}{\sqrt{2}} & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & \frac{1}{\sqrt{2}} & \frac{1}{\sqrt{2}} & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & \frac{1}{\sqrt{2}} & \frac{1}{\sqrt{2}} \\ \frac{1}{2} & \frac{1}{2} & -\frac{1}{2} & -\frac{1}{2} & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & \frac{1}{2} & \frac{1}{2} & -\frac{1}{2} & -\frac{1}{2} \\ \frac{1}{2} & -\frac{1}{2} & \frac{1}{2} & -\frac{1}{2} & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & \frac{1}{2} & -\frac{1}{2} & \frac{1}{2} & -\frac{1}{2} \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ x_3 \\ x_4 \\ x_5 \\ x_6 \\ x_7 \\ x_8 \end{bmatrix}. \tag{2.28}$$

The equation illustrates the difference between the wavelet matrix $\mathbf{H}_{DWT}$ and the wavelet packet matrix $\mathbf{H}_{WP}$. The rows are changed according to the determined WP decomposition.

2.5 Gabor Frames (GF)

In this section, we firstly briefly explain the Gabor frames (GF) transform in general [32], [39]. Then, an implementation of the GF transform based on a generalised DFT filter bank with a symmetric signal extension is presented leading to a matrix representation of the GF transform.

2.5.1 Gabor Frames Theory

To introduce Gabor Frames (GF) we start with an explanation of the sampled Short Time Fourier Transformation (STFT), [40]. For the continuous case, the STFT for a function $x(\tau)$ equals:

$$(STFT_g x)(\omega, t) = \int x(\tau) g(t-\tau) e^{-jc\omega\tau} d\tau, \tag{2.29}$$

where $g(t-\tau)$ is a window function that specifies a certain window type (e.g. rectangular) and restricts $x(\tau)$ to a certain time interval. This is the reason for the term "short time" Fourier transformation which addresses that not the whole spread of $x(\tau)$ in the time domain is transformed by one step t. The formula for the inverse transformation with the above specified window function is

$$x(\tau) = \frac{1}{2\pi ||g||^2} \int \int (STFT_g x)(\omega, t) g^*(t-\tau) e^{jc\omega\tau} d\omega dt \tag{2.30}$$

In the following, we will regard the modulation term $e^{-jc\omega\tau}$ as part of the window function. If $\{g_{j,k}\}$ is a dictionary of time-frequency shifted versions (indexed by j and k as for the DWT) of a single window function g, it is called a Gabor frame [41]. A Gabor frame has the additional property that there exist constants $A, B > 0$, so-called frame bounds such that

$$A \parallel x \parallel^2 \leq \sum_{j,k} \in \mathbb{Z} \parallel < x[n], g_{j,k}[n] > \parallel^2 \leq B \parallel x \parallel^2 \quad \forall x \in H, \tag{2.31}$$

where H is the Hilbert space of functions $x(\tau)$. This property guarantees the completeness of $g_{j,k}$ meaning that any signal $x \in H$ can be represented as an absolutely convergent infinite series of the $g_{j,k}$ or in the finite case, a linear combination thereof.

The discrete Gabor Frames transformation [42] becomes:

$$(G_g x)[j,k] = c_{j,k} = \sum_{n\in\mathbb{Z}} x[n] g_{j,k}[n] = \sum_{n\in\mathbb{Z}} x[n] g_j[k \cdot D - n] \qquad j,k,D \in \mathbb{Z}, \tag{2.32}$$

where the $c_{j,k}$ are called Gabor coefficients and $g[n]$ is addressed as a Gabor elementary function. The parameter D denotes the time sampling interval. The $g_{j,k}[n]$ follow the equation

$$g_{j,k}[n] = g[k \cdot D - n] \cdot e^{-2\pi jcjb_G n}. \tag{2.33}$$

The $g_{j,k}[n]$ are time-shifted and modulated copies of the elementary function $g[n]$ where b_G denotes the frequency sampling interval. The formula for the inverse transformation [43] is

$$x[n] = \sum_{j,k\in\mathbb{Z}} c_{j,k} g_{j,k}[n]. \tag{2.34}$$

The coefficients $c_{j,k}$ span a lattice in the time frequency plane as it is shown in Figure 2.14. We

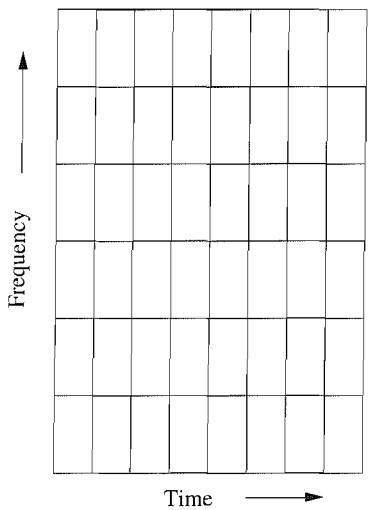

Figure 2.14: Time-frequency tiling for Gabor Frames (GF).

see that the area that is covered by one coefficient $c_{j,k}$ is a rectangle. Like the DWT the GF is a fixed transformation. However, by choosing elementary functions that have a different time-frequency tiling (changing the length and width of the rectangles), one elementary function can be selected that parameterises the data best.

The DWT is based on a signal representation with basis functions. The most important distinction between basis functions and frames is that the transformation coefficients in the basis representation are unique. But the transformation coefficients in the frame representation need not to be unique. For example in (2.34), we do not require the $g_{j,k}$ to be orthogonal, nor the $c_{j,k}$ to be unique. The frames provide signal representations with more freedom.

For our application of the GF transform, we need to ensure that the transform matrix $\mathbf{H}$ according to (2.1) is unitary to result in a fixed energy relation for the transform as it is the case for the DWT and WP. For this we require the transform to be based on a tight frame which is defined as having equal frame bounds $A = B$. Therefore, if the transformation equation equals $\mathbf{y} = \mathbf{H}_{GF}\mathbf{x}$, the vectors will show the following property:

$$\frac{\| \mathbf{y} \|^2}{\| \mathbf{x} \|^2} = c, \tag{2.35}$$

with c being a constant. To state this explicitly: When we speak of GF transform from now on, we always require the frames to have equal constant frame bounds.

2.5.2 GF Based on a GDFT Filter Bank

The GF transform can be implemented via an oversampled generalised DFT (GDFT) filter bank, see e.g. [44], [45] or [46], which fulfils the condition of possessing a tight frame meaning equal frame bounds. A general structure of a filter bank is shown in Figure 2.15. The bank decomposes a signal

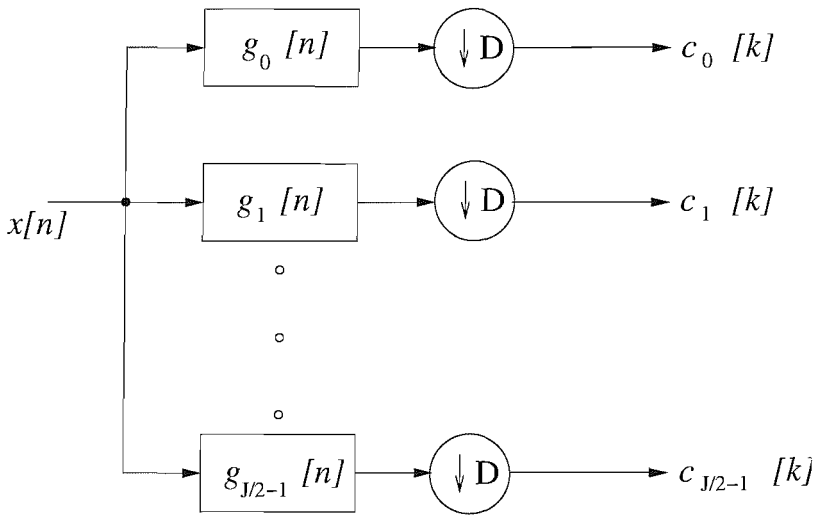

Figure 2.15: GDFT filter bank with J-channels and decimation ratio D.

$x[n]$ into J subbands, each produced by a branch $g_j[n]$ by convolution and decimation by a factor D resulting in the Gabor coefficients $c_j[k]$. The convolution and the down-sampling represent the time shift indexed by k, the index j accounts for different frequency bands equalling the frequency shift for the Gabor transform. The analysis filters $g_j[n]$ are derived from a real valued prototype FIR filter $p[n]$ of even length N_F by a GDFT,

$$g_j[n] = t_j[n] \cdot p[n], \qquad t_j[n] = e^{-jc\frac{2\pi}{J}(j+j_{off})(n+n_{off})}, \qquad n, j \in \mathbb{N}. \tag{2.36}$$

The term generalised DFT stems from offsets j_{off} and n_{off} introduced into the frequency and time indices [44]. The transform that is conducted by the filter bank is complex valued and as our data

is real, it is sufficient to cover the frequency range $0 < f < f_{a/2}$ where $f_{a/2}$ denotes half of the sampling frequency because the skipped frequency range contains the same information of the data as it is conjugated complex. As an example the magnitude responses of an 8-channel filter bank are presented in Figure 2.16. We see that the Gabor coefficients $c_{j,k}$ are obtained by filtering with the

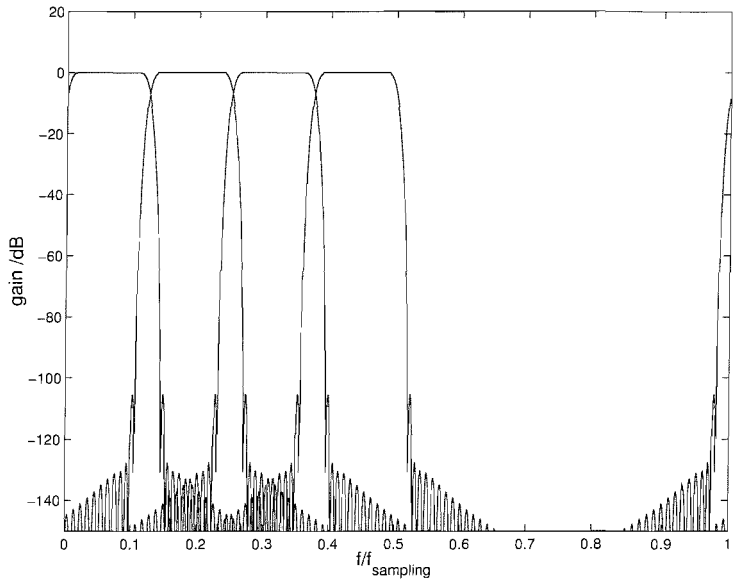

Figure 2.16: Spectra of a 8-channel GDFT filter bank where the conjugated complex component is skipped for real data input.

modulated window $g_j[n]$ which is the modulated prototype filter $p[n]$ describing the j-th channel. The time-shift is accomplished by convolution (index k). This assumes a continuous data stream; to deal with data on a finite interval, again an appropriate signal extension or border filter is required.

2.5.3 GF Transformation Matrix $\mathbf{H}_{GF}$

Based on [47], we implement the GF transform over a GDFT filter bank with symmetric and even length filters.

As the GDFT filter bank is oversampled [44], redundancy is introduced into the transformed signal. This increases the length for each frequency band compared to the critically sampled case. The amount of redundancy is determined by the decimation ratio D. If the discrete function $x[n]$ in Figure 2.15 is represented by a vector $\mathbf{x}$ with length N and a GDFT filter bank with J frequency bands and with a symmetric extension is applied, the length K_{FB} for each frequency band will be by

$$K_{FB} = N/D + 1 \qquad D < J, \tag{2.37}$$

where the "+1" indicates the symmetric extension according to Figure 2.3 for an even length filter. Also, N needs to be an integer multiple of D.

As we analyse real data only, half of the frequency range can be skipped because it contains the same information as it is conjugated complex. Hence, by considering just half of the frequency bands we arrive at a vector length K for the transform vector $\mathbf{y}$ obtained by GF transform with

$$K = K_{FB} \cdot J/2 = (N/D + 1) \cdot J/2. \tag{2.38}$$

As mentioned above, N needs to be an integer multiple of D. In practise, this condition can be enforced by truncating the data in $\mathbf{x}$. Although truncation can be potentially harmful to data, in the later application to evoked emissions, for example the pre-stimulus interval does not carry any vital information and therefore offers leeway for adjustment of the data segment $\mathbf{x}$.

The transformation matrix $\mathbf{H}_{GF}$ is complex valued and contains in its rows the modulated and translated impulse responses of a prototype low pass filter. According to $\mathbf{y} = \mathbf{H}_{GF}\mathbf{x}$, $\mathbf{y}$ contains the results of the convolutions of the input signal with the respective filters for the different frequency bands.

In the following we show a general procedure of how to obtain the matrix $\mathbf{H}_{GF}$.

The signal to be transformed is given by: $\mathbf{x} \in \mathbb{C}^N$. The filter vector $\mathbf{h} \in \mathbb{C}^{N_F}$ represents the impulse response of one modulated window function $g_j[n]$. The decimation ratio is D and the filter bank to be implemented has J subbands in total. We aim at a transformation for one subband j according to $\mathbf{y}_j = \mathbf{H}_j\mathbf{x}$, with $\mathbf{y}_j \in \mathbb{C}^{K_{FB}}$. K_{FB} follows (2.37). As mentioned in the previous sections, we show how to implement the symmetric extension of $\mathbf{x}$. For this, we define a symmetrically extended version $\tilde{\mathbf{x}}$ of $\mathbf{x}$ over an extension matrix $\mathbf{E}$ for one subband according to

$$\tilde{\mathbf{x}} = \mathbf{E} \cdot \mathbf{x} \tag{2.39}$$

with $\mathbf{E} \in \mathbb{Z}^{N+N_F \times N}$. The matrix $\mathbf{E}$ has the structure:

$$\mathbf{E} = \begin{bmatrix} \mathbf{A} \\ \mathbf{I} \\ \mathbf{A}^F \end{bmatrix}, \tag{2.40}$$

with $\mathbf{A} \in \mathbb{Z}^{N_F/2 \times N}$. The matrix $\mathbf{I}$ is a $N \times N$ identity matrix. The matrix $\mathbf{A}^F$ is a "flip" matrix of $\mathbf{A}$ defined as:

$$\mathbf{A}^F = \mathbf{J}^{R\times R} \cdot \mathbf{A} \cdot \mathbf{J}^{N\times N}, \tag{2.41}$$

with R being the number of rows of $\mathbf{A}$ and $\mathbf{J}$ being a reverse identity matrix defined as

$$\mathbf{J} = \begin{bmatrix} & & & 1 \\ & & \cdot & \\ & \cdot & & \\ 1 & & & \end{bmatrix}. \tag{2.42}$$

The structure of the matrix $\mathbf{A}$ depends on N_F and N. The following condition describes the relation between these parameters:

$$2sN < N_F \leq 2(s+1)N$$

with $s \in \mathbb{N}^0$. We can differentiate two cases:

- s is even or 0;
- s is odd.

For the first case, $\mathbf{A}$ is constructed as follows:

$$\mathbf{A} = \begin{bmatrix} \mathbf{J}^{N_F/2-sN \times N_F/2-sN} & \mathbf{0}^{N_F/2-sN \times (s+1)N-N_F/2} \\ & \mathbf{I} \\ & \mathbf{J} \\ & \mathbf{I} \\ & \mathbf{J} \\ & \cdot \\ & \cdot \\ & \cdot \end{bmatrix} \Bigg\} s \tag{2.43}$$

with the matrix $\mathbf{0}$ containing only zeros.

For the second case, $\mathbf{A}$ equals:

$$\mathbf{A} = \begin{bmatrix} \mathbf{0}^{N_F/2-sN \times (s+1)N-N_F/2} & \mathbf{I}^{N_F/2-sN \times N_F/2-sN} \\ & \mathbf{J} \\ & \mathbf{I} \\ & \mathbf{J} \\ & \cdot \\ & \cdot \\ & \cdot \end{bmatrix} \Bigg\} s \tag{2.44}$$

This shows the construction of $\tilde{\mathbf{x}}$ as a suitably extended vector of $\mathbf{x}$. Now, we need to define the filtering of $\tilde{\mathbf{x}}$ with the GDFT filter bank:

$$\mathbf{y}_j = \tilde{\mathbf{H}}_j \cdot \tilde{\mathbf{x}}$$

with

$$\tilde{\mathbf{H}}_j = \begin{bmatrix} \mathbf{h}^T & & & & \\ \mathbf{o}_D^T & \mathbf{h}^T & & & \\ \mathbf{o}_D^T & \mathbf{o}_D^T & \mathbf{h}^T & & \\ & & & \ddots & \\ & & & & \mathbf{h}^T \end{bmatrix}, \tag{2.45}$$

where $\tilde{\mathbf{H}}_j \in \mathbb{C}^{K_{FB} \times N_F+N}$. The vector $\mathbf{o}_D$ has D elements which are zeros.

Now we can formulate the GF decomposition for one subband:

$$\mathbf{y}_j = \tilde{\mathbf{H}}_j \cdot \mathbf{E} \cdot \mathbf{x} = \mathbf{H}_j \cdot \mathbf{x} \tag{2.46}$$

with $\mathbf{H}_j \in \mathbb{C}^{K_{FB} \times N}$. Consequently, we arrive at the GF decomposition for $\mathbf{x}$ as:

$$\mathbf{y} = \begin{bmatrix} \mathbf{H}_1 \\ \cdot \\ \cdot \\ \cdot \\ \mathbf{H}_{J/2} \end{bmatrix} \cdot \mathbf{x} = \mathbf{H}_{GF} \cdot \mathbf{x}, \tag{2.47}$$

with $\mathbf{y} \in \mathbb{C}^K$.

Finally, to illustrate these statements we show an example of the transformation matrix $\mathbf{H}_{GF}$ for a symmetrically extended signal vector $\mathbf{x}$. Let us consider the case that the length for $\mathbf{x}$ is $N = 12$ and assume that a complex filter bank with $J = 4$ frequency bands and decimation ratio $D = 3$ is applied. The length of the filter vector $\mathbf{h}$ shall be equal to the length of the signal $\mathbf{x}$ ($N = N_F$). Then, by using (2.38), the length for $\mathbf{y}$ equals $K = 10$. The matrix $\tilde{\mathbf{H}}_j$ equals

$$\tilde{\mathbf{H}}_j^{5\times 24} = \begin{bmatrix} \mathbf{h}^T & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & \mathbf{h}^T & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & \mathbf{h}^T & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & \mathbf{h}^T & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & \mathbf{h}^T \end{bmatrix}$$

according to (2.45). The matrix $\mathbf{E}$ that conducts the symmetric signal extension for all channels is

$$\mathbf{E}^{24\times 12} = \begin{bmatrix} \mathbf{J}^{6\times 6} & \mathbf{0}^{6\times 6} \\ \mathbf{I}^{12\times 12} & \\ \mathbf{0}^{6\times 6} & \mathbf{J}^{6\times 6} \end{bmatrix}.$$

Then, $\mathbf{H}_j$ is obtained by multiplication of the matrices

$$\mathbf{H}_j^{5\times 12} = \tilde{\mathbf{H}}_j^{5\times 24} \cdot \mathbf{E}^{24\times 12}.$$

Hence, $\mathbf{H}_j$ has the following structure

$$\mathbf{H}_j^{5\times 12} =$$

$$\begin{bmatrix} h_6+h_7 & h_5+h_8 & h_4+h_9 & h_3+h_{10} & h_2+h_{11} & h_1+h_{12} & 0 & 0 & 0 & 0 & 0 & 0 \\ h_3+h_4 & h_2+h_5 & h_1+h_6 & h_7 & h_8 & h_9 & h_{10} & h_{11} & h_{12} & 0 & 0 & 0 \\ h_1 & h_2 & h_3 & h_4 & h_5 & h_6 & h_7 & h_8 & h_9 & h_{10} & h_{11} & h_{12} \\ 0 & 0 & 0 & h_1 & h_2 & h_3 & h_4 & h_5 & h_6 & h_7+h_{12} & h_8+h_{11} & h_9+h_{10} \\ 0 & 0 & 0 & 0 & 0 & 0 & h_1+h_{12} & h_2+h_{11} & h_3+h_{10} & h_4+h_9 & h_5+h_8 & h_6+h_7 \end{bmatrix}.$$

Consequently, for our case $\mathbf{y}$ equals

$$\mathbf{y} = \mathbf{H}_{GF} \cdot \mathbf{x} = \begin{bmatrix} \mathbf{H}_{subband\ 1} \\ \mathbf{H}_{subband\ 2} \end{bmatrix} \cdot x = \begin{bmatrix} y1_{subband\ 1} \\ \cdot \\ \cdot \\ \cdot \\ y5_{subband\ 1} \\ y1_{subband\ 2} \\ \cdot \\ \cdot \\ \cdot \\ y5_{subband\ 2} \end{bmatrix}.$$

This basic principle is used to attain the matrix $\mathbf{H}_{GF}$ that analyses signals with longer support later for the application. Equally to DWT and WP, one transformation coefficient for the GF can be interpreted as a least squares fit between the modulated and translated elementary Gabor functions and the signal $\mathbf{x}$. Again, $\mathbf{y}$ shows how good an elementary Gabor function matches $\mathbf{x}$ (see Figure 2.5)

or we can say that $\mathbf{y}$ indicates when a certain frequency appears in $\mathbf{x}$ (see Figure 2.14). The only difference is that the elementary Gabor functions are not orthogonal; they are linearly dependent. Because the matrix $\mathbf{H}_{GF}$ contains only half of the frequency subbands, $\text{rank}\{\mathbf{H}_{GF}\} = K$. This means that the rows of $\mathbf{H}_{GF}$ are linearly independent. In other words, the comparison of the GF with orthogonal transforms (as illustrated in Figure 2.5) yields, that the elementary Gabor functions can have different angles among them and there are more than needed to describe the signal to be analysed which introduces redundancy.

Apart from the restriction that N needs to be an integer multiple of D, there are no constraints in terms of the length of the signal to be analysed and the length of the analysing filter. This is a clear advantage of the presented GF transform compared to other approaches.

2.6 Summary

We have presented a unifying representation for various popular transforms in matrix notation, where a symmetric extension (extension by reflection, [26]) of the data is incorporated into $\mathbf{H}$. It is not always the numerically most efficient way of computation. However, while fast implementations of DWT, WP [30] and GF [48] avoid matrix implementations, the calculation of a limited number of significant elements in $\mathbf{y}$ can be performed faster by extracting the according rows from $\mathbf{H}$.

Chapter 3

TF Transform Based Feature Selection

This chapter deals with feature selection methods for TF transformed data, as shown in Figure 1.1 in the middle. We start with Section 3.1 by explaining receiver operating (ROC) analysis, as it can be used to evaluate the separability performance of the succeeding feature selection methods. Section 3.2 gives a simple energy reduction approach to select TF coefficients that contain most of the signal energy. If we deal with data, where only fewer than 60 measurements for one TF coefficient are available, Section 3.3 states statistical tests can be used to choose coefficients that show a statistical significant difference. Section 3.4 follows showing how TF coefficients containing a difference can be selected to increase the overall separability. This is done by introducing a SNR-like criterion that calculates a similarity measure for two partial averages of measured data. The separability applying the criterion is increased by an ROC analysis based selection of TF coefficients. Therefore, the ROC analysis is used twice: Firstly, for finding a selection of TF coefficients and secondly, for the evaluation of the found TF coefficients. The chapter concludes with a summary in Section 3.5.

3.1 ROC Analysis

A good measure for differentiation between two distributions are ROC curves [49], since the area under the ROC curve measures the separability independent of the selection of any threshold. Therefore, they have become remarkably useful in medical decision-making [50]. As they are very substantial to our work, they are described in more detail next.

Firstly, we start by introducing the terms sensitivity and specificity [50]. We assume we a have population consisting of healthy controls and patients that suffer from a certain disease but do not know or cannot express their suffering (e.g. hearing loss in infants). Our goal is to determine the patient group out of the population. For this, we run an imaginary test on the population. The outcome of the test consists of one test parameter which is either positive indicating the tested person is diseased or negative meaning the tested person is healthy. In order to evaluate the performance of that test, the following values can be used, as illustrated in Table 3.1.

The interrelationship equations in the table result from the fact that each person is classified as healthy or diseased by the test or in other words that a decision is made. In the following the terms represented in the table will be used equivalently, meaning that we always refer to the true positive

	Test for disease		interrelationship
	diseased group	healthy group	
Test result: positive	true positive (TP) rate in %, sensitivity, hit rate	false positive (FP) rate in %, false alarm rate	TP + FN = 100%
Test result: negative	false negative (FN) rate in %	true negative (TN) rate in %, specificity	TN + FP = 100%

Table 3.1: Definition of sensitivity and specificity.

rate when speaking of sensitivity or hit rate.

Secondly, we define predictive values. The sensitivity and specificity indicate the performance of test whether it is useful for diagnosis or not. They do not give the information that e.g. a positive test means a person is diseased with a certain probability based on the performance of the test. This information is given by the predictive values. The positive predictive value (PPV) is defined as the proportion of people that are diagnosed as positive and are actually diseased:

$$\text{PPV} = \frac{\text{TP}}{\text{TP+FP}}. \tag{3.1}$$

Respectively, the negative predictive value (NPV) equals:

$$\text{NPV} = \frac{\text{TN}}{\text{TN+FN}}. \tag{3.2}$$

Thirdly, we introduce the ROC curves. An ROC curve is a graphical representation of the trade off between sensitivity and specificity for every possible cut off. By tradition, the plot of the ROC curve shows the false positive rate on the x axis and the hit rate on the y axis. However, based on the interrelationships shown in Table 3.1, the axis of the ROC curve can be modified. Suppose the above mentioned test parameter yields distributions for the diseased and healthy groups as illustrated in Figure 3.1 on the left.

The solid line represents the distribution of the test parameter for the patient group, the dashed line for healthy controls on the left in the figure. In the upper case the distributions have a distance of their means equalling two standard deviations, $d = 2$ whereby the distributions possess the same variances. The upper right of the figure shows the resulting ROC curve. The lower case shows the same but for $d = 0.1$. In order to explain the meaning and the difference of the area under the ROC curve which is given by F in the Figure 3.1, Table 3.2 shows sample values taken from the figure as indicated by the circles in Figure 3.1 (right). Table 3.2 uses the same layout as Table 3.1 to state the TP, FP, FN and TN rates.

Ideally, for a good separation, the sensitivity and the specificity should be very high. As Table 3.2 illustrates, a value for the area under the ROC curve close to 1 yields a relatively good separation, whereas a value close to 0.5 yields a very poor performance when taking both the sensitivity and specificity into account.

Having introduced the ROC analysis, we continue by introducing the feature selection methods that can be evaluated upon their separability performance by the area under the ROC curve.

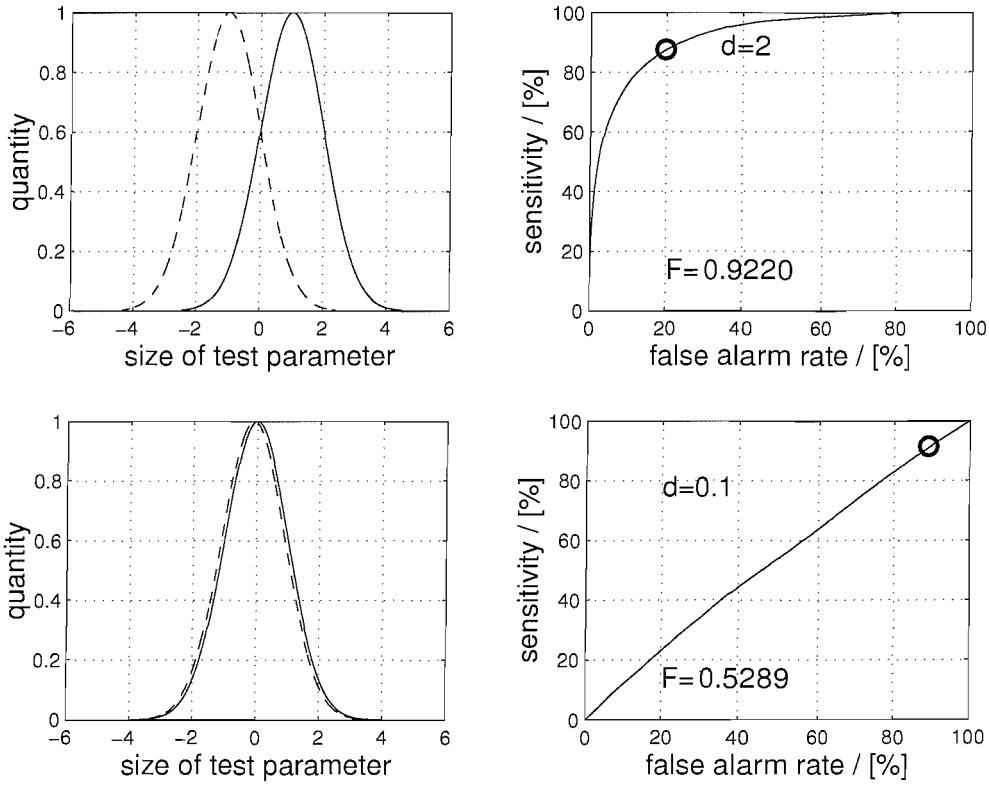

Figure 3.1: ROC explanation: sample distributions (left) for diseased (solid) and healthy (dashed) groups assuming an imaginary test parameter yielding ROC curves (right). The circles indicate the example values shown in Table 3.2

	Approximated sample values upper case in Figure 3.1		Approximated sample values lower case in Figure 3.1	
	"solid" group	"dashed" group	"solid" group	"dashed" group
positive test result	90%	20 %,	90%	90%
negative test result	10%	80 %	10%	10%

Table 3.2: Example values represented by the circles in Figure 3.1.

3.2 Energy Reduction

To select the features from TF transformed data, a sensible and promising approach is to reduce the energy of the transformed data. The energy of the transformed vector $\mathbf{y}$ according to (2.1) equals:

$$E = \sum_{k=0}^{K-1} y_k^2 \tag{3.3}$$

with y_k being the elements of the vector $\mathbf{y}$ with dimension K.

An energy reduction can be conducted by setting the smallest TF coefficients whose sum-of-squares makes up for a certain amount of the total signal power to zero. Given the spares nature of the TF transforms, the proportion of such coefficients relative to the total number of coefficients is much greater than the energy reduction proportion, hence reducing computational effort at subsequent

stages.

The complete approach is explained by an example. Suppose a discrete signal is given by the vector $\mathbf{x}$ as shown in Figure 3.2 top. The signal is corrupted by noise referred to as $\mathbf{x}_{\text{noisy}}$.

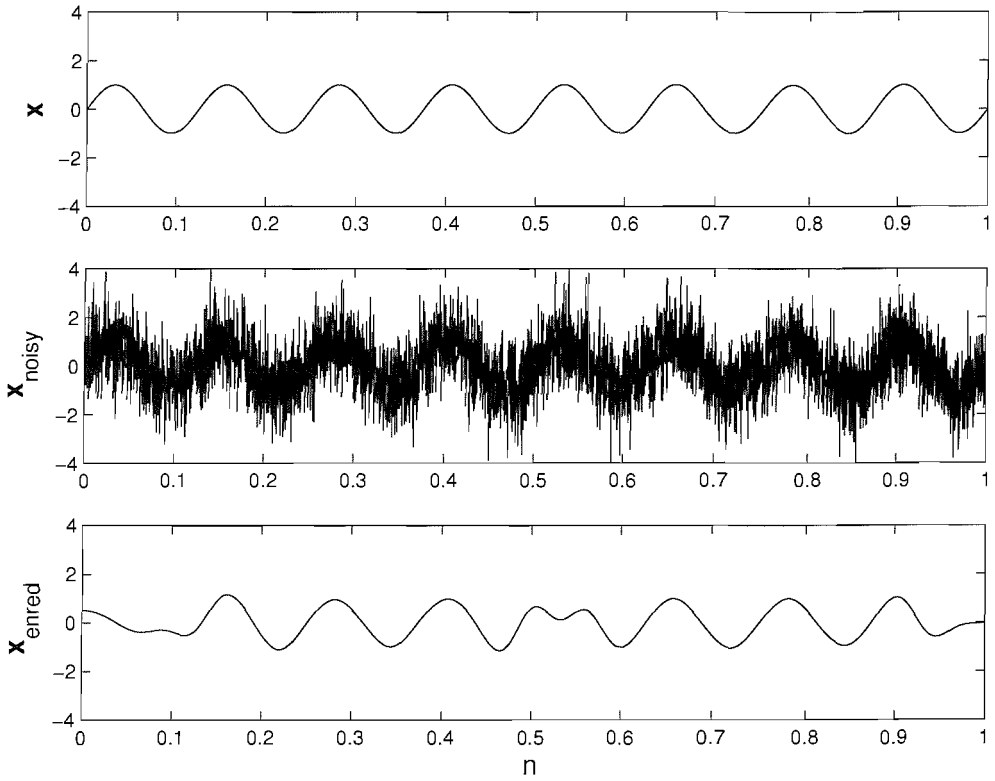

Figure 3.2: Example for energy reduction: (top) the signal $\mathbf{x}$ is corrupted by noise (middle). Discarding DWT coefficients that contribute less than 50% to the total energy yields the signal $\mathbf{x}_{\text{enred}}$ (bottom).

To reduce the noise, the vector is transformed according to (2.1) with a DWT applying Mallat's wavelet [30] for which good results concerning the analysis of biomedical data have been reported [13]. Figure 3.3 shows a basis function derived from the Mallat wavelet where the impulse response corresponds to the $25th$ row of $\mathbf{H}_{DWT}$ for a signal length of $N = 1024$ for the signal $\mathbf{x}_{\text{noisy}}$ to be analysed. It is a scaled and translated version of the mother wavelet. The figure also shows the magnitude response of this wavelet illustrating its band-pass characteristic or the selected detail in the frequency domain. The frequency f_s is the sampling frequency.

Now, we apply an energy reduction to the transformed data vector $\mathbf{y}$ whereby all coefficients whose sum-of-squares contribute less than 50% to total energy are discarded. This yields only 13 out of $K = 1024$ for the example illustrated in Figure 3.2. Multiplying the reduced transformed vector with $\mathbf{H}_{DWT}^{-1}$ yields $\mathbf{x}_{\text{enred}}$ illustrated in Figure 3.2 (bottom).

We see that the original sine wave is relatively well reconstructed and that the TF transform based energy reduction denoised the signal significantly revealing its main characteristic. Note that the relatively bad reconstruction in the middle is coincidental and is not connected to a characteristic of the transform. Concluding, it can be said that this approach seems to be well suited to select features for a later diagnosis or classification. Moreover, if we applied a diagnostic test assuming we

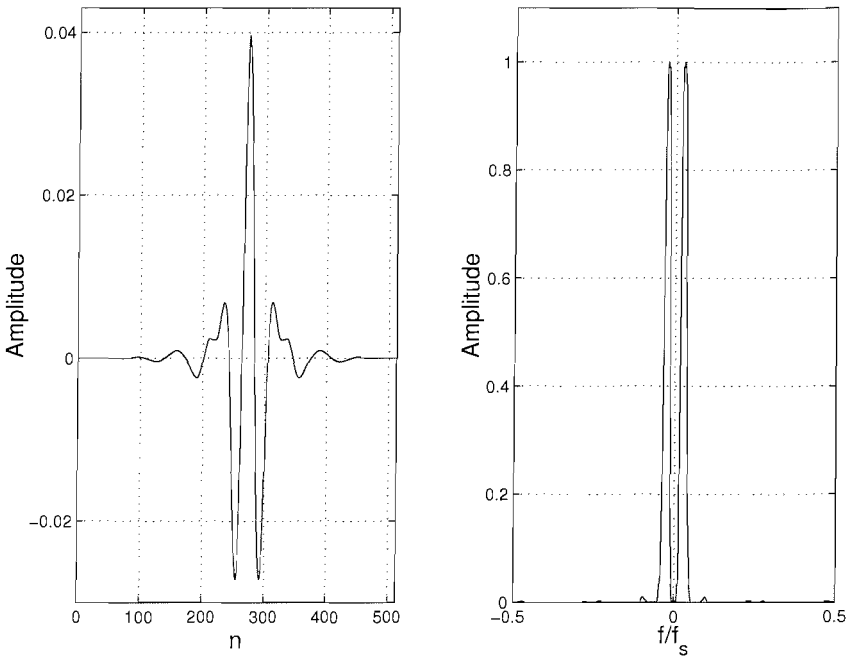

Figure 3.3: Basis function derived from the Mallat wavelet: (left) impulse response and (right) magnitude response.

have got data from healthy and diseased people, we could conduct a ROC analysis for each selected TF coefficient yielding the area under the ROC curve which we could use to evaluate the separability performance of each TF coefficient.

To test the obtained TF coefficients by the energy reduction in terms of their statistical significance, especially when only few measurements are available, statistical tests can be applied which is shown next.

3.3 Statistical Tests

Based on an energy reduction of TF transforms coefficients, we want to identify coefficients that allow us to conduct a statistical significant differentiation for two data sets with respect to their mean value when only few measurements are available. For this, the following statistical tests can be used.

3.3.1 F-test

Prior to the selection of significant coefficients that represent the main characteristics of the data, an F-test [51] is conducted to determine which method is used to identify them. The aim of this test is to determine whether two data sets are sampled from normal distributions with the same variances. If a value for the significance level P of lower than 0.05 is obtained by the F-test, we conclude that the hypothesis is rejected and the two data sets are sampled from normal distributions having different variances. The value of $P = 0.05$ is a limit commonly used in medical research [51]. When the sets $\mathbf{x}_1$ and $\mathbf{x}_2$ contain the series for one time or transformed coefficient for all measurements indexed by l taken for two data sets that ought to be separated, they can be compared by the F-value, which is

given by [51]

$$F_v = \frac{\sigma_1^2}{\sigma_2^2}, \tag{3.4}$$

with σ_1^2 and σ_2^2 being the variances of the two data sets. To find the significance level P for the F-test, we need to define the degrees of freedom for the two data sets according to

$$\begin{aligned} \nu_1 &= L_1 - 1 \text{ and} \\ \nu_2 &= L_2 - 1, \end{aligned} \tag{3.5}$$

with L_1 and L_2 being the number of samples, ν_1 the degrees of freedom for the data set 1 and ν_2 the degrees of freedom for the data set 2. With the F-value defined by (3.4) and the degrees of freedom ν_1 and ν_2, the significance level P for the F-test can be determined from lookup tables in the literature, e.g. [51]. The tabulated F-values are all greater than 1, thus the two data sets in (3.4) need to be labelled such that $\sigma_1^2 \geq \sigma_2^2$. If the outcome of the F-test confirms that the two data sets are sampled from distributions with equal variances, we can subsequently conduct a t-test to determine distinctive coefficients. If the result of the F-test is that the underlying distributions from which the two data groups are sampled possess different variances we conduct a ut-test. The t-test and the ut-test are defined in the next subsection.

3.3.2 t- and ut-Tests

The t-test gives the probability that two data sets sampled from potentially two different distributions with identical variance possess different mean values, for which a significance is returned. The t-value is defined as [51]

$$t_v = \frac{\overline{\mathbf{x}}_1 - \overline{\mathbf{x}}_2}{\sqrt{\frac{\sigma_1^2}{L_1} + \frac{\sigma_2^2}{L_2}}} = \frac{\overline{\mathbf{x}}_1 - \overline{\mathbf{x}}_2}{\sigma\sqrt{\frac{1}{L_1} + \frac{1}{L_2}}} \quad , \tag{3.6}$$

with $\sigma^2 = \sigma_1^2 = \sigma_2^2$. The values $\overline{\mathbf{x}}_1$ and $\overline{\mathbf{x}}_2$ represent the means for the two data sets, according to

$$\overline{\mathbf{x}}_i = \frac{1}{L_i} \cdot \sum_{l=0}^{L_i-1} x_i[l], \qquad i \in \{1, 2\}, \tag{3.7}$$

with $\mathbf{x}_i^T = [x_i[0]\ x_i[1]\ ...\ x_i[L_i - 1]]$.

The t-value also corresponds to a certain significance level P, which can be looked up from tables [51], with the degrees of freedom defined by $\nu_t = \nu_1 + \nu_2 = L_1 + L_2 - 2$. A smaller value for P indicates that the data sets have a significantly different mean. For example, for $P = 0.01$ the probability that the differences in the means are due to a sampling error is 1%. To identify distinctive coefficients, the determination of the applied significance level will be discussed in the next subsection. The two tested distributions are the distributions for a specific transform parameter over the two data sets.

For the case that the F-test yields a difference in variances such that the t-test cannot be used, we apply a ut-test for unequal variances defined as

$$ut = \frac{\overline{\mathbf{x}}_1 - \overline{\mathbf{x}}_2}{\sqrt{\frac{\sigma_1^2}{L_1} + \frac{\sigma_2^2}{L_2}}} \quad . \tag{3.8}$$

According to [51], for data sets sampled from distributions with unequal variances, the t-distribution can be approximated by the ut value if the t-table is entered at the following defined degree of freedom:

$$\nu_{ut} = \frac{(\sigma_1^2/L_1 + \sigma_2^2/L_2)^2}{\frac{(\sigma_1^2/L_1)^2}{L_1-1} + \frac{(\sigma_2^2/L_2)^2}{L_2-1}}. \tag{3.9}$$

This test tends to be less powerful than the usual t-test, since it uses fewer assumptions [51]. The main purpose of the ut-test is to have an analysis tool for all coefficients at hand whether they show equal variances or not.

To determine a significance level P, the relation of the t-test to the receiver operating characteristic (ROC) analysis is shown in the next subsection.

3.3.3 ROC Back Test

Here, we make use of the ROC analysis to evaluate and obtain a significance level for the t-tests or ut-tests. The relation is investigated as follows. Different values for the area under the ROC curve are determined. Then, two Gaussian distributions are generated with zero mean, a standard deviation of one and a sample size of 100000 for each distribution. Then, a constant is added to each sample of one distribution to change its mean. That constant is adjusted so that a ROC analysis of the two distributions equals the previously determined ROC curve value. From these distributions, a certain number of random samples are taken out and based on a t-test or ut-test, the significance level for the t-test is calculated for these samples originating from the Gaussian distributions. This calculation is repeated with random samples from the distributions and the significance level is averaged until it converges. The ROC analysis is independent of the sample size whereas the t-test and ut-test depend on it. Therefore, different sample sizes yield different relations, which is illustrated in Figure 3.4 for a significance level P found by the t-test. In more detail, to obtain one point in the curve, the corresponding number of samples are taken out of the distributions which possess the illustrated area under the ROC curve. Then, a significance level P is calculated by the t-test as shown in the previous subsection . This procedure is repeated until the significance level of the t-test converges to the value shown in the figure. When using the ut-test to investigate this relation, the results are very similar, e.g. no differences can be observed when also including the resulting curves on the plot in Figure 3.4. Note that the same number of samples is used for the ut-test as for the t-test.

At this point, we emphasise that the studied relation between the area under the ROC curve value and a significance level P of a t-test is based on Gaussian distributions with different means but with a standard deviation of one each. The area under the ROC curve represents a parameter which evaluates both the distance between the means and different standard deviations for two distributions. We considered the variance of the distributions as constant and changed the difference in the means observing the effect on the area under the ROC curve value.

To illustrate the above statements, let us take the analysis of the panic disorder in Chapter 5, which deals with a sample size of 24. Table 3.3 shows the relation between the areas under the ROC curve and the most commonly used significance levels P [51] for a t-test resulting from our back test approach, in more detail.

In most social research significance levels of $P = 0.05$ or $P = 0.01$ are used to determine difference

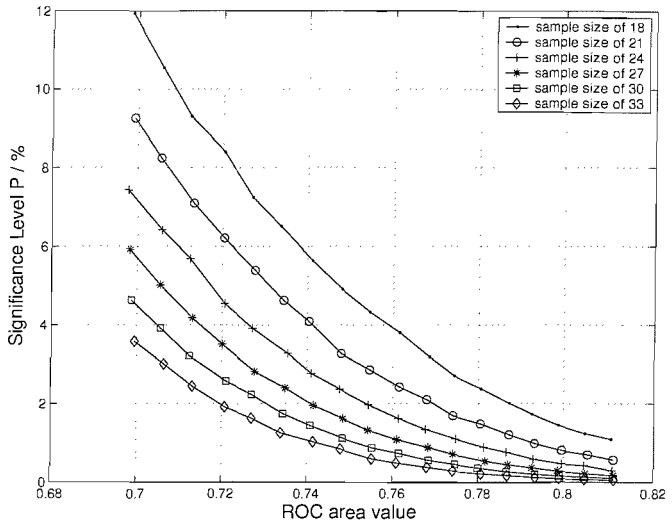

Figure 3.4: Significance Level P for t-test over area under the ROC curve value for different sample sizes.

Area under the ROC curve	Significance level P of a t-test according to our back test approach
0.717	0.05
0.778	0.01

Table 3.3: Area under ROC curve and significance levels P for a sample size of 24.

between two sets of data [51]. In other studies such as [13], ROC values of ≈ 0.77 are found and stated to yield acceptable separation performance. Therefore, we choose a significance level of $P = 0.01$ for our studies to obtain distinctive transform coefficients. However, we are aware that this selection is based on the assumption that the distributions characterised by a ROC value of 0.778 possess a standard deviation of one each. As we used this method only to determine a certain significance level for the t-test or back test results yielded by a t-test with a ROC analysis, this assumptions seems justified.

The introduced methods allow us to select a certain amount of TF coefficients that can be used for diagnosis. However, they are found on the basis of containing a certain energy or a significant difference in their means. To select TF coefficients within the ones selected by the energy reduction method on the basis of their ability to increase the separability value obtained by the area under the ROC when added to one coefficient set, a SNR-like criterion is introduced in the next section.

3.4 SNR-like Criterion

So far, only TF coefficients containing much energy have been selected as features, which by means of statistical tests can also be applied when only little data is available. Among the coefficients obtained by the energy reduction method, we aim to select the ones that contain a significant difference, for diagnosis.

3.4.1 Difference Evaluation for Differential Diagnosis

Let a measured time series of a biomedical signal be represented by a vector $\mathbf{x}$. For one subject, repeated measurements that contain the response to a certain stimulus can be placed into a data matrix according to

$$\mathbf{D} = \begin{bmatrix} \mathbf{x}^T[1] \\ \mathbf{x}^T[2] \\ \vdots \\ \mathbf{x}^T[L] \end{bmatrix} = [\mathbf{sp}[1]\ \mathbf{sp}[2]\ \cdots\ \mathbf{sp}[N]] \tag{3.10}$$

with $\mathbf{sp}[n]$ being the columns of the data matrix containing the distribution of the samples for the responses to the stimulus with length L. As biomedical data is prone to a considerable contamination by noise, measurement equipment, e.g. [52] performs averaging over a total of L responses $\mathbf{x}[l]$, $l = 0(1)L-1$, whereby two partial averages can be labelled with subscripts A and B:

$$\overline{\mathbf{x}}_{\mathrm{A}} = \frac{2}{L}\sum_{l=0}^{L/2-1} \mathbf{x}[2l], \tag{3.11}$$

$$\overline{\mathbf{x}}_{\mathrm{B}} = \frac{2}{L}\sum_{l=0}^{L/2-1} \mathbf{x}[2l+1]. \tag{3.12}$$

Thus, our transformation equation that conducts the parameterisation becomes:

$$\mathbf{y}_i = \mathbf{H}_{j_T} \cdot \overline{\mathbf{x}}_i \quad , \tag{3.13}$$

where $\mathbf{y}_i$ is a vector holding the transformation coefficients, $i = \{\mathrm{A}, \mathrm{B}\}$ selects a partial average, e.g. according to (3.11) or (3.12) and $j_T = \{\mathrm{DWT}, \mathrm{WP}, \mathrm{GP}\}$ the potential transform method.

Standard assessment of biomedical data, e.g. transient evoked otoacoustic emissions [52] considers the correlation between the two partial averages, i.e. the value $\rho = \overline{\mathbf{x}}_{\mathrm{A}}^{\mathrm{T}} \cdot \overline{\mathbf{x}}_{\mathrm{B}}$, or an SNR value

$$\mathrm{SNR} = \frac{\|\overline{\mathbf{x}}_{\mathrm{A}} + \overline{\mathbf{x}}_{\mathrm{B}}\|_2^2}{\|\overline{\mathbf{x}}_{\mathrm{A}} - \overline{\mathbf{x}}_{\mathrm{B}}\|_2^2 + \gamma} \quad , \tag{3.14}$$

where γ is a small constant to avoid division by zero. If signal information is contained in both $\mathbf{x}_{\mathrm{A}}$ and $\mathbf{x}_{\mathrm{B}}$, it will be subtracted out in the denominator leaving noise only; in the numerator, the signal information will be add up, yielding a sum of signal information plus noise from both averages. Therefore, this measure is described as a SNR-like ratio. These SNR values or the above mentioned correlation can be compared to a given threshold, yielding a binary decision for the presence or absence of biomedical signal information. Further tests to detect signal information according to [6] are a variance ratio F_r and a modified variance ratio F_{SP}^*:

$$F_r = \frac{Var(\overline{\mathbf{x}}_{\mathrm{A}} + \overline{\mathbf{x}}_{\mathrm{B}})}{Var(\overline{\mathbf{x}}_{\mathrm{A}} - \overline{\mathbf{x}}_{\mathrm{B}})} \tag{3.15}$$

and

$$F_{SP}^* = \frac{Var(\frac{\overline{\mathbf{x}}_{\mathrm{A}} + \overline{\mathbf{x}}_{\mathrm{B}}}{2})}{\frac{1}{N}\sum_{n=0}^{N-1} Var(\mathbf{sp}[n])} \tag{3.16}$$

with $\mathbf{sp}[n]$ being the columns of a data matrix defined by (3.10) and Var being the variance. According to [6], especially the F_{SP}^* test is more powerful than the tests implemented in known screening devices, namely the correlation coefficient and the SNR value. This shows that there is space for improving the known SNR value, at which we aim in the following.

3.4.1.1 Modified SNR Criterion

As the partial averages $\overline{\mathbf{x}}_i$ according to (3.13) are finite, a symmetric extension of the data is incorporated into $\mathbf{H}_{j_T}$ according to Section 2.2. The SNR criterion in (3.14) can be applied similarly to the transform vectors, since the Euclidean vector norm is invariant under orthonormal matrices $\mathbf{H}_{j_T}$ [28], which also holds for a GF transform matrix with tight frame bounds [42],

$$\begin{aligned} \xi_{\mathrm{SNR}} &= \frac{\|\overline{\mathbf{x}}_{\mathrm{A}} + \overline{\mathbf{x}}_{\mathrm{B}}\|_2^2}{\|\overline{\mathbf{x}}_{\mathrm{A}} - \overline{\mathbf{x}}_{\mathrm{B}}\|_2^2 + \gamma} = \frac{\|\mathbf{H}_{j_T}(\overline{\mathbf{x}}_{\mathrm{A}} + \overline{\mathbf{x}}_{\mathrm{B}})\|_2^2}{\|\mathbf{H}_{j_T}(\overline{\mathbf{x}}_{\mathrm{A}} - \overline{\mathbf{x}}_{\mathrm{B}})\|_2^2 + \gamma} \\ &= \frac{\|\mathbf{y}_{\mathrm{A}} + \mathbf{y}_{\mathrm{B}}\|_2^2}{\|\mathbf{y}_{\mathrm{A}} - \mathbf{y}_{\mathrm{B}}\|_2^2 + \gamma} \quad . \end{aligned} \tag{3.17}$$

For further consideration, (3.17) can be written in element-wise notation,

$$\xi_{\mathrm{SNR}} = \frac{\sum_k |y_{\mathrm{A}}[k] + y_{\mathrm{B}}[k]|^2}{\sum_k |y_{\mathrm{A}}[k] - y_{\mathrm{B}}[k]|^2 + \gamma} \quad . \tag{3.18}$$

Differences in the distribution of this criterion between different data groups will be employed for differentiation.

Let us assume that the coefficients containing information of the biomedical data to be studied are given in the transform domain as

$$y_i[k] = \begin{cases} c_{\mathrm{data}}[k] + s_i[k] & , \quad k \in C_{\mathrm{opt}} \\ s_i[k] & , \quad k \in C_{\mathrm{opt}}^{\perp} \end{cases} , \tag{3.19}$$

where only the set C_{opt} includes coefficients with a significant contribution $c_{\mathrm{data}}[k]$, and $s_i[k]$ is a random variable with zero mean and variance σ^2. We further assume independence between $s_{\mathrm{A}}[k]$ and $s_{\mathrm{B}}[k]$ $\forall k$. With the expectation operator $\mathcal{E}\{\cdot\}$, we have

$$\mathcal{E}\left\{ \sum_{k \in C_{\mathrm{opt}}^{\perp}} |y_{\mathrm{A}}[k] \pm y_{\mathrm{B}}[k]|^2 \right\} = 2 \sum_{k \in C_{\mathrm{opt}}^{\perp}} \sigma^2 .$$

Considering the SNR criterion in the form

$$\xi_{\mathrm{SNR}} = \frac{\sum_{k \in C_{\mathrm{opt}}} |y_{\mathrm{A}}[k] + y_{\mathrm{B}}[k]|^2 + \sum_{k \in C_{\mathrm{opt}}^{\perp}} |y_{\mathrm{A}}[k] + y_{\mathrm{B}}[k]|^2}{\sum_{k \in C_{\mathrm{opt}}} |y_{\mathrm{A}}[k] - y_{\mathrm{B}}[k]|^2 + \sum_{k \in C_{\mathrm{opt}}^{\perp}} |y_{\mathrm{A}}[k] - y_{\mathrm{B}}[k]|^2 + \gamma} \quad , \tag{3.20}$$

it is evident that, on average, the summations over set $C_{\mathrm{opt}}^{\perp}$ add the same constant term in numerator and denominator. Therefore, a stronger criterion for absence or presence of $c_{\mathrm{data}}[k]$ is constructed according to

$$\hat{\xi}_{\mathrm{SNR}} = \frac{\sum_{k \in C_{\mathrm{opt}}} |y_{\mathrm{A}}[k] + y_{\mathrm{B}}[k]|^2}{\sum_{k \in C_{\mathrm{opt}}} |y_{\mathrm{A}}[k] - y_{\mathrm{B}}[k]|^2 + \gamma} \tag{3.21}$$

by omitting the coefficient set $C_{\mathrm{opt}}^{\perp}$. This shows that, by aiming at a good parameterisation of the data, an improved criterion $\hat{\xi}$ over (3.20) can be attained.

The criterion defined by (3.21) contains a fundamental difference to the correlation coefficient ρ. For the parameterised partial averages, this coefficient can be expanded as

$$\rho = \overline{\mathbf{x}}_{\mathrm{A}}^H \cdot \overline{\mathbf{x}}_{\mathrm{B}} = \mathbf{y}_{\mathrm{A}}^H \cdot \mathbf{H}_{j_T}^{\dagger H} \cdot \mathbf{H}_{j_T}^{\dagger} \cdot \mathbf{y}_{\mathrm{B}} = \mathbf{y}_{\mathrm{A}}^H \cdot \mathbf{y}_{\mathrm{B}}, \tag{3.22}$$

since $\mathbf{H}_{j_T}^{\dagger^H} \cdot \mathbf{H}_{j_T}^{\dagger} = \mathbf{I}$ with $\mathbf{I}$ being the unitary matrix.

By using (3.19), ρ becomes

$$\rho = \sum_{k \in C_{\text{opt}}} (c_{\text{data}}[k] + s_{\text{A}}[k])(c_{\text{data}}[k] + s_{\text{B}}[k]) + \sum_{k \in C_{\text{opt}}^{\perp}} (s_{\text{A}}[k] s_{\text{B}}[k]).$$

Hence, the expectation value for ρ equals

$$\mathcal{E}\{\rho\} = \sum_{k \in C_{\text{opt}}} c_{\text{data}}[k]^2, \tag{3.23}$$

because the noise is uncorrelated to itself and coefficients that contain signal information of the data. As the noise averages out, it is not necessary to select C_{opt} when using the correlation coefficient ρ for assessment of data information. It would possess the same distinctiveness as if all coefficients were chosen, meaning that a feature selection approach based on the correlation coefficient would yield the same separability results as applying no feature selection at all.

3.4.1.2 Differentiating Groups

To distinguish between two different data groups, the criterion $\hat{\xi}_{\text{SNR}}$ has to yield distinct, separable distributions for both groups. A good measure for this separability are receiver operating characteristic (ROC) curves [49], since the area under the ROC curve measures the separability independent of the selection of any threshold as introduced in Section 3.1.

The coefficient set C_{opt} on which the evaluation of $\hat{\xi}_{\text{SNR}}$ will be based may differ from the one described by (3.21). Of the coefficients that are indicative of the groups' signal information, only those representing a difference between groups rather than the mere presence or absence of signal information will be considered. Nevertheless, the proposed improvement of the criterion in 3.4.1.1, through disposing of non-indicative coefficients remains valid.

3.4.1.3 Set Selection

To determine the set C_{opt} for the differentiation of two data groups, an iterative approach similar to [13] is proposed in the following. A starting value is found by evaluating the separability based on the ROC area of $\hat{\xi}_{\text{SNR}}[k]$ for every single transform coefficient with index $k = 0(1)511$. Then from a suitably selected starting coefficient, the set C_{opt} is grown by stepwise inclusion of further coefficients that maximise the separability of $\hat{\xi}_{\text{SNR}}$ until no further improvement is gained.

For the starting value, the separability of $\hat{\xi}_{\text{SNR}}[k]$ for each single transform coefficient is measured over the training data containing two different data groups. From these coefficients, in [13] the one yielding the highest separability (denoted as first maximum) is chosen as a seed for C_{opt}. Alternatively, for reasons detailed below, we will further consider that the second largest coefficient (denoted as second maximum) may be selected instead. Note that simply choosing the coefficients whose individual values $\hat{\xi}_{\text{SNR}}[k]$ separate well does not yield the optimal set C_{opt}. Therefore, an iterative growth of C_{opt} is suggested, starting with a set C_0 containing the first selected coefficient only.

The iterative growth at step i from set C_{i-1} to C_i includes one additional coefficient. The index r of this coefficient is determined by evaluating the separability of the criterion at step i,

$$\hat{\xi}_{\mathrm{SNR}}^{(i)}[r] = \frac{(y_{\mathrm{A}}[r]+y_{\mathrm{B}}[r])^2+\sum_{k\in C_{i-1}}(y_{\mathrm{A}}[k]+y_{\mathrm{B}}[k])^2}{(y_{\mathrm{A}}[r]-y_{\mathrm{B}}[r])^2+\sum_{k\in C_{i-1}}(y_{\mathrm{A}}[k]-y_{\mathrm{B}}[k])^2+\gamma}, \tag{3.24}$$

over all possible $r \in C_{i-1}^{\perp}$. The coefficient maximising the separability of $\hat{\xi}_{\mathrm{SNR}}^{(i)}$ is added to the coefficient set, $C_i = C_{i-1} \cup r$.

To narrow the complexity of the algorithm, the search can be restricted to a neighbourhood of the coefficients contained in C_{i-1}. We consider neighbourhood of first order as direct adjacency of coefficients in the TF plane as performed in [13]. As a modification of this, and to provide a slightly larger search area, second order neighbourhood additionally includes all coefficients that are directly adjacent to first order neighbours. An additional benefit besides the reduced search complexity is the potentially improved generalisation to other data if adjacency is favoured over random coefficient placement, since the latter is easier prone to model noise in the data.

The iteration is finally stopped if the separability of $\hat{\xi}_{\mathrm{SNR}}^{(i)}$ is not increased over $\hat{\xi}_{\mathrm{SNR}}^{(i-1)}$. Then the set of coefficients that are considered as significant for a differentiation between two hearing ability groups is complete, $C_{\mathrm{opt}} = C_{i-1}$.

By this difference evaluation method, we aim at identifying differences in the TF characteristic of two data groups. If the TF characteristic of one data group fully overlaps the TF coefficients of the second data group, this method covers all possible areas, which can be used for differentiation of the two data groups. But if the second data group is not fully covered by the coefficients of data group one, this characteristic can also be used to increase the separability which is discussed in detail next.

3.4.2 General Difference Evaluation Method for Data Groups with Partially Disjoint Features

Before we propose a generalised approach to Subsection 3.4.1, we motivate this extension by an example. Let us assume we apply the difference evaluation method described in the previous subsection to data groups showing a TF characteristic as illustrated in Figure 3.5 according to (3.19), where the black blocks represent constant coefficients c_{data} and the noise s_i is zero.

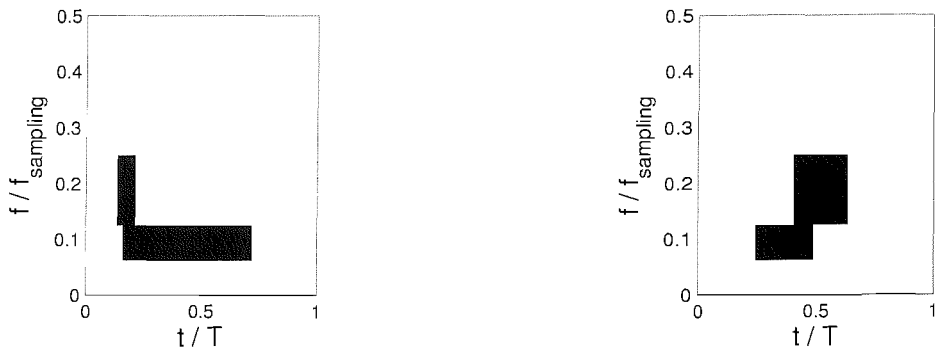

Figure 3.5: Sample TF coefficient distribution for data group one (left) and data group two (right).

We aim at separating these two data groups as well as possible with the difference evaluation method described in Subsection 3.4.1. By deploying this method, we aim at identifying the coefficient set illustrated in Figure 3.6 left, when the data groups are contaminated with noise. The reason

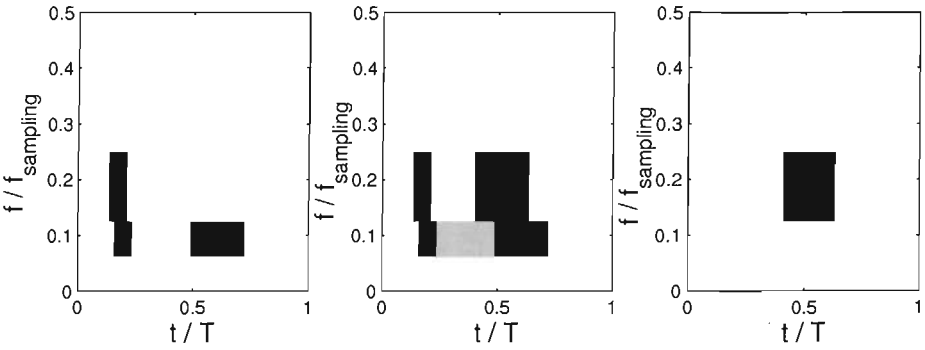

Figure 3.6: (Left) coefficient set C^1_{opt} originating from data group one, (middle) coefficients (grey) that do not represent a difference, and (right) coefficient set C^2_{opt} originating from data group two.

that the difference evaluation method according to Subsection 3.4.1 results in identifying only the coefficients C^1_{opt} originating from data group one is that these represent a difference to data group two. The grey coefficients shown in the middle part of Figure 3.6 do not contain a difference for the two data groups and are therefore not useful for increasing the separability for the two groups. The coefficients representing a difference originating from data group two are not identified as the applied search method aims at maximising the separability of the criterion according to (3.24). Therefore, the following question arises: How can we identify the coefficients originating from data group two that contain a difference to data group one (Figure 3.6 right)? Furthermore, how can we use this coefficient set C^2_{opt} originating from data group two to increase the overall separability by defining a criterion for the whole coefficient set C_{opt}?

To address the first question, we have a closer look at the ROC area analysis. Based on a TF characteristic for the transform coefficients $\mathbf{y}_i$ shown in Figure 3.5, Figure 3.7 shows ROC area values of for every single transform coefficient $\mathbf{y}$. For this sample analysis, 50 data vectors according to (3.19) for each $\mathbf{y}_A$ and $\mathbf{y}_B$ for each data group were generated with a relatively small noise contamination. For these noisy data vectors, the ROC area values for the distributions $\xi_{\text{SNR}}[k]^1$ and $\xi_{\text{SNR}}[k]^2$ were calculated.

Figure 3.7 illustrates the coefficients shown in Figure 3.6 (left) with their ROC area value being close to one. The coefficients coming from data group two show a ROC area value close to zero, that can be observed in Figure 3.6 (right). The grey coefficients in Figure 3.6 middle do not show ROC area values that indicate a difference.

In the following, we explain why the ROC analysis results in values close to zero and not just between 0.5 and one. Figure 3.8 shows the calculation of the ROC area value for 2 Gaussian distributions. Let us assume, that (left) the solid line is the distribution of ξ^1_{SNR} for a coefficient k of data group one, and the dotted line the distribution of ξ^2_{SNR} for the same coefficient k of data group two. If the analysed distribution of ξ^1_{SNR} contains signal information of the data group one, $k \in C^1_{\text{opt}}$, we will expect it to have a positive mean. If the analysed distribution of ξ^2_{SNR} contains noise only for

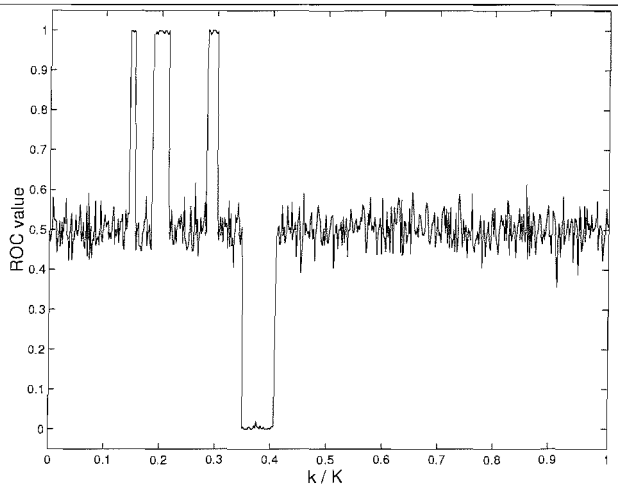

Figure 3.7: ROC area values for sample data groups with very little noise contamination.

data group two, $k \in C_{\text{opt}}^{\perp}$ we will expect it to have a mean close to zero. The figure (top) illustrates this case for two distributions with a distance of their means equalling the standard deviation d. This explains how the values close to one in Figure 3.7 are attained. The figure (top right) also shows the run of the ROC curve for two distributions completely covering each other, $d = 0$ (dotted line).

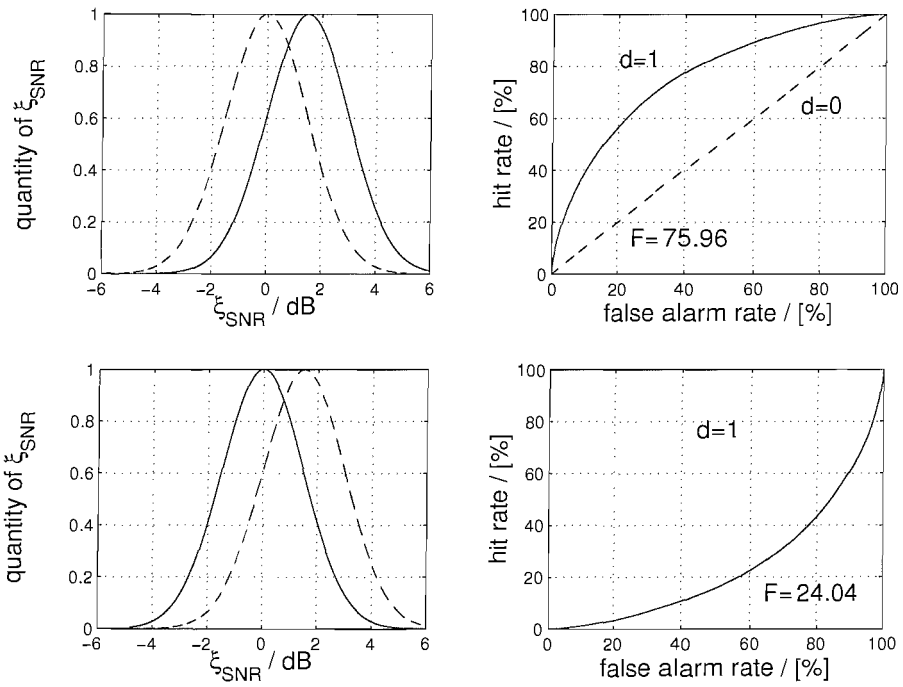

Figure 3.8: (Left) ξ_{SNR} distribution for (solid) data group one and (dotted) data group two, and (right) ROC curves.

The bottom part of the figure shows the case for the distribution ξ_{SNR}^2 of a coefficient $k \in C_{\text{opt}}^2$ of data group two, analysed with the same coefficient k of data group one and for this group $k \in C_{\text{opt}}^{\perp}$. This leads to a ROC area value that is below 0.5 as illustrated in the figure (bottom right). This case corresponds to the values close to zero in Figure 3.7.

With these explanations the question can be formulated as follows: We compare two groups of data, represented by $\overline{\mathbf{x}}^1_{\{\mathrm{A,B}\}}$ and $\overline{\mathbf{x}}^2_{\{\mathrm{A,B}\}}$. For some coefficients k the criterion or measure according to (3.18) yields an average of $\overline{\xi}^1_{SNR}[k] > \overline{\xi}^2_{SNR}[k]$ and for others, $\overline{\xi}^1_{SNR}[k] < \overline{\xi}^2_{SNR}[k]$. Both cases should be exploited to yield a good discrimination.

The background of this problem is that we do not know which distribution is the reference. Usually, for a ROC analysis the distributions are known, there is no reference required. For our case, it is important to known from which distribution to start and in which direction the threshold is shifted during the ROC analysis. We could also just flip the sign of the analysis meaning we would look for the ROC values obtained by one minus the ROC values shown in Figure 3.7. However, for illustration purposes and a formulation of a combined SNR-criterion the discussion as it will be shown next seems more sensible.

3.4.2.1 Reciprocal Difference Evaluation Method

Based on the above explanations, the answer to the question how we can identify the coefficient set C^2_{opt} originating from data group two can be given. We define a reciprocal criterion according to

$$\hat{\xi}^{\mathrm{rez}}_{\mathrm{SNR}} = \frac{\sum_{k\in C^2_{\mathrm{opt}}} |y_{\mathrm{A}}[k] - y_{\mathrm{B}}[k]|^2}{\sum_{k\in C^2_{\mathrm{opt}}} |y_{\mathrm{A}}[k] + y_{\mathrm{B}}[k]|^2 + \gamma}, \tag{3.25}$$

with the distinctive coefficient set for data group two referred to as C^2_{opt} from now on. To explicitly state this, the coefficient set C^1_{opt} refers to coefficients originating from data group one, yielding a difference to their counterparts of data group two which are an element of $C^{\perp}_{\mathrm{opt}}$. The coefficient set C^2_{opt} refers to coefficients originating from data group two, yielding a difference to their counterparts of data group one which are an element of $C^{\perp}_{\mathrm{opt}}$.

The ROC analysis based on this reciprocal criterion corresponds to the case is shown in Figure 3.9. Note that this figure refers to the reciprocal criterion and not Figure 3.8. Similar as before, we assume that (left) the solid line is the distribution of $\xi^{\mathrm{rez}\,1}_{\mathrm{SNR}}$ for a coefficient k of data group one, and the dotted line the distribution of $\xi^{\mathrm{rez}\,2}_{\mathrm{SNR}}$ for the same coefficient k of data group two. If the analysed distribution of $\xi^{\mathrm{rez2}}_{\mathrm{SNR}}$ contains signal information of the data group two, $k \in C^2_{\mathrm{opt}}$, we will expect it have a negative mean this time. If the analysed distribution of $\xi^{\mathrm{rez1}}_{\mathrm{SNR}}$ contains noise only for data group one, $k \in C^{\perp}_{\mathrm{opt}}$ we will expect it to have a mean close to zero. Figure 3.9 (top) illustrates this case for two distributions with a distance of their means equalling the standard deviation d.

We see that with the reciprocal criterion, the ROC area analysis results in identifying a distinctive coefficient of data group two. However, as before, the bottom part of Figure 3.9 shows the case for the distribution $\xi^{\mathrm{rez1}}_{\mathrm{SNR}}$ of a coefficient $k \in C^1_{\mathrm{opt}}$ of data group one, analysed with the same coefficient k of data group two and for this group $k \in C^{\perp}_{\mathrm{opt}}$. This leads to a ROC area value that is below 0.5 as illustrated in the figure (bottom right).

Based on the reciprocal criterion according to (3.25), the ROC area values shown in Figure 3.7 would be reversed. The values originating from data group one which are close to one would be close to zero, and the values originating from data group two which are close to zero would be close to one.

To identify the coefficient set C^2_{opt} that maximises the separability, we just grow the coefficient set

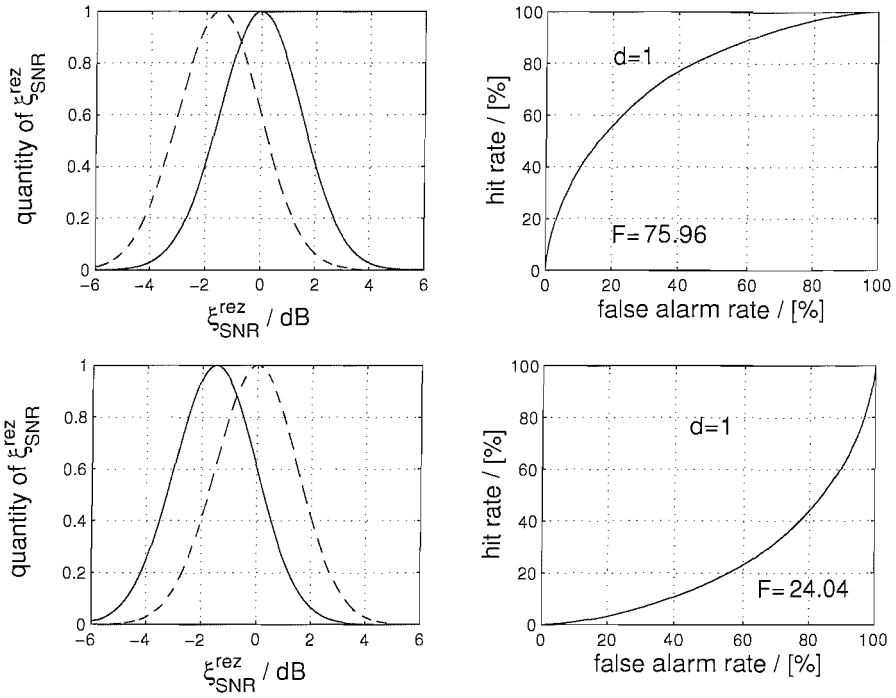

Figure 3.9: (Left) $\xi^{\text{rez}}_{\text{SNR}}$ distribution for (solid) data group one and (dotted) data group two, and (right) ROC curves.

according to (3.24) to maximise the separability, however applying to the following reciprocal equation

$$\hat{\xi}^{\text{rez}\ (i)}_{\text{SNR}}[r] = \frac{(y_{\text{A}}[r]-y_{\text{B}}[r])^2+\sum_{k\in C^2_{i-1}}(y_{\text{A}}[k]-y_{\text{B}}[k])^2}{(y_{\text{A}}[r]+y_{\text{B}}[r])^2+\sum_{k\in C^2_{i-1}}(y_{\text{A}}[k]+y_{\text{B}}[k])^2+\gamma},$$

yielding the estimated coefficient set C^2_{opt}.

Now, having stated that by deploying iterative searches according to (3.24) and (3.26) we yield coefficient sets C^1_{opt} and C^2_{opt} that contain the difference between two data sets, we will answer the question how a combined criterion can be developed that uses the combined coefficient set $C_{\text{opt}} = C^1_{\text{opt}} \cup C^2_{\text{opt}}$.

3.4.2.2 Combining Coefficient Sets C^1_{opt} and C^2_{opt} to C_{opt} to Maximise Separability

Based on the criteria (3.21) and (3.25) we propose two combined criteria in the following. In the next Subsection, we will evaluate and compare their performance for constructed artificial data.

Firstly, a sum-like combination of (3.21) and (3.25) yields:

$$\hat{\xi}^{+}_{\text{SNR}} = \frac{(\sum_{k\in C^1_{\text{opt}}}|y_{\text{A}}[k]+y_{\text{B}}[k]|^2)+(\sum_{k\in C^2_{\text{opt}}}|y_{\text{A}}[k]-y_{\text{B}}[k]|^2)}{(\sum_{k\in C^1_{\text{opt}}}|y_{\text{A}}[k]-y_{\text{B}}[k]|^2)+(\sum_{k\in C^2_{\text{opt}}}|y_{\text{A}}[k]+y_{\text{B}}[k]|^2)+\gamma}. \tag{3.26}$$

The idea for this proposal is the following: According to (3.19) for a coefficient $k \in C^1_{\text{opt}}$, the numerator will yield the sum of the energy of the signal information plus twice the noise. The denominator will contain twice the noise only. For $k \in C^2_{\text{opt}}$, the denominator contains the signal plus twice the noise

energy, the numerator the twice the noise energy only. This is expected to result in an improved separability compared to the results obtained by the single criteria according to (3.21) or (3.25).

The second proposal is a product of (3.21) and (3.25):

$$\hat{\xi}^{*}_{\mathrm{SNR}} = \frac{(\sum_{k\in C^1_{\mathrm{opt}}} |y_{\mathrm{A}}[k] + y_{\mathrm{B}}[k]|^2) \cdot (\sum_{k\in C^2_{\mathrm{opt}}} |y_{\mathrm{A}}[k] - y_{\mathrm{B}}[k]|^2)}{(\sum_{k\in C^1_{\mathrm{opt}}} |y_{\mathrm{A}}[k] - y_{\mathrm{B}}[k]|^2) \cdot (\sum_{k\in C^2_{\mathrm{opt}}} |y_{\mathrm{A}}[k] + y_{\mathrm{B}}[k]|^2) + \gamma}. \tag{3.27}$$

For a coefficient $k \in C^1_{\mathrm{opt}}$ the numerator contains a product of the signal information plus noise originating from $k \in C^1_{\mathrm{opt}}$, times one noise component originating from the second data group. The denominator equals the product for two noise components. For $k \in C^2_{\mathrm{opt}}$, this statement is reversed meaning we yield a fraction of noise times noise divided by noise times the sum of noise and signal energy. Therefore, it can be expected that this criterion performs worse than (3.26) as the noise originating from a coefficient set containing no signal information enters the fraction as a product and no just a sum.

To test and confirm the introduced methods for differential diagnosis, the next subsection discusses and shows the separability results for artificial data groups.

3.4.3 Feature Selection for Constructed Data

In this subsection, we generate artificial vectors $\mathbf{y}_i$ to test and validate the difference evaluation method to separate two data groups.

3.4.3.1 Artificial Data Groups

Exemplary in this subsection, the synthetic coefficients $\mathbf{y}_i$ are based on a DWT parameterisation, j_T = DWT, see Figure 2.6, as the application of the difference evaluation method is independent of the parameterisation method.

Our synthetic data group one is illustrated in Figure 3.10 left by the coefficients $\mathbf{y}_i$ that are contained by C_{opt}. The presented time TF plane is based on DWT TF tiling and vector length of $\mathbf{y}_i$

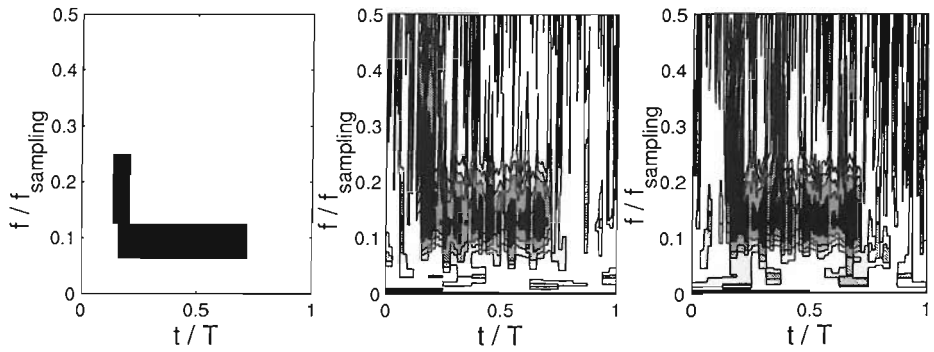

Figure 3.10: Created artificial data group one: (left) $\mathbf{y}_i$ without noise, (middle) sample $\mathbf{y}_{\mathrm{A}}$ contaminated by noise, and (right) sample $\mathbf{y}_{\mathrm{B}}$ contaminated by noise with t/T being the normalised time axis.

of $K = 512$. The parts in the middle and on the right of the figure show a contamination of the left

part by the same amount of noise. Real biomedical data can reveal a similar TF characteristic. When we conduct simulations to evaluate the performance of our difference evaluation method next, we will increase the amount of noise stepwise. To conduct a differential diagnosis, we need to determine a second data group, which is shown in Figure 3.11.

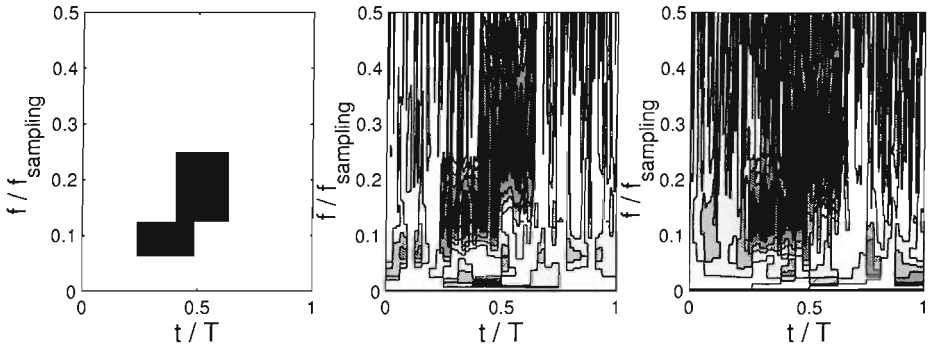

Figure 3.11: Created artificial data group two: (left) $\mathbf{y}_i$ without noise, (middle) sample $\mathbf{y}_\mathrm{A}$ contaminated by noise, and (right) sample $\mathbf{y}_\mathrm{B}$ contaminated by noise with $\mathbf{t}/\mathrm{T}$ being the normalised time axis.

3.4.3.2 Simulation Adjustment

To apply the methods discussed in the previous subsections, two data sets, $\mathbf{y}_\mathrm{A}$ and $\mathbf{y}_\mathrm{B}$ are created for each data group, based on a TF characteristic shown above. For each $\mathbf{y}_\mathrm{A}$ and $\mathbf{y}_\mathrm{B}$ for each data group, 50 data sets are generated with a length of $K = 512$ (corresponding to 50 subjects for each data group for a real life case). The coefficients, that are illustrated by Figures 3.10 and 3.11 on the left, are set to a constant value, and then a certain amount of noise is added, according to

$$\mathrm{SNR}_\mathrm{EN} = 10 \cdot \log[\frac{\frac{K_C}{K} \cdot c_\mathrm{data}^2}{\sigma^2}] \text{ dB} \tag{3.28}$$

with SNR_EN denotes the signal to noise ratio for energies, K_C is the number of constants c_data and σ^2 the variance of the Gaussian noise according to (3.19), dB refers to the unit Decibel and log is the logarithm with the basis 10. Based on this SNR_EN, we create artificial data groups from -18 dB to -7 dB with a step-size varying form 1 dB to 0.5 dB. Figures 3.10 and 3.11 in the middle and on the right show one data set out of the 50 for $\mathbf{y}_\mathrm{A}$ and $\mathbf{y}_\mathrm{B}$ for $\mathrm{SNR}_\mathrm{EN} = -13.5$ dB.

When we apply the method of difference evaluation to the artificial data groups, we obtain an estimated coefficient set C_opt^1 originating from data group one and C_opt^2 from data group two for the criteria defined by (3.26) and (3.27). E.g, Figure 3.12 shows these coefficient sets separately for illustration purposes for $\mathrm{SNR}_\mathrm{EN} = -13.5$ dB. The combined separability when using these coefficient sets for the ROC area analysis according to (3.26) and (3.27) results in 0.8864 and 0.8677, respectively.

To confirm and back-test our results, control data groups are created. For the identified coefficient sets, the separability for these control data groups is calculated. The result of this test may give some hints whether the difference evaluation method is adapted to the data groups, form which they were generated or whether the method is generally valid.

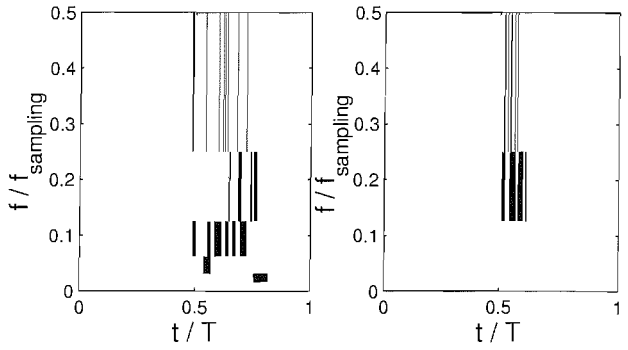

Figure 3.12: Estimated coefficient set (left) C^1_{opt} for data group one and (right) C^2_{opt} for data group two with a separability value of ≈ 0.82 when regarding each case separately.

As mentioned in 3.4.1.3, for identifying distinctive coefficients, the method is restricted to search only in first and second order neighbourhood of coefficients that have already been identified as significant. To justify this approach, simulations for an exhaustive search were also conducted. Exhaustive search means, that the stepwise growth of the distinctive coefficient set is not limited, meaning that all coefficients are tested for increasing the separability by one iteration step. The expectation is that the exhaustive search shows better separability results but performs very badly for the control data and is hence not generalisable. To illustrate this search, Figure 3.13 shows the estimated coefficient set C_{opt} separately for C^1_{opt} and C^2_{opt} for illustration purposes, where the combined separability results in the maximum separability value, 1.0000 and 1.0000 for $\hat{\xi}^+$ and $\hat{\xi}^*$ respectively.

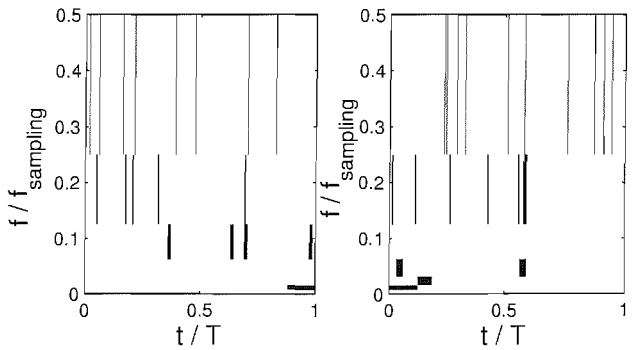

Figure 3.13: Estimated coefficient set (left) C^1_{opt} for data group one with a separability value of 0.9996, and (right) C^2_{opt} for data group two with a separability value of 0.9980 for an exhaustive search.

Based on these statements, we finally present the simulation results for the general difference evaluation method for artificial data.

3.4.3.3 Simulation Results and Discussion

In Figure 3.14, the simulation results are shown for first and second order neighbourhood coefficient growth compared for different criteria, including a test for control data groups. The results confirm the expectation, the criterion $\hat{\xi}^+$ yields better results than $\hat{\xi}^*$, for both the data used for adjustment and the control groups. Also, the application of both criteria is justified, as they result in improved separabilities compared the criteria aiming to identify the uncombined coefficient sets C^1_{opt} or C^2_{opt}.

In Figure 3.15 the results for the above mentioned exhaustive search are stated. The expectation is confirmed that this search yields better separability results but performs poorer for the control data groups meaning that the generalisation is better ensured by growing the coefficient set only by the first and second order neighbourhood. To underline this statement, Figure 3.16 illustrates the performance of the sum-like criterion according to (3.26) for the control data groups for a first and second order neighbourhood search compared with an exhaustive search.

Based on these results, we conclude that the difference evaluation method to conduct a differential diagnosis of artificial data yields reasonable separability results. Hence, we apply the method to real life data parameterised by linear transformations.

3.5 Summary

In this chapter we have shown how features among TF transformed data can be selected. We firstly introduced ROC analysis which can be used as an evaluation method for our approaches. Then, an energy reduction of the transformed data was shown underlined by an example. For only little data, statistical tests were explained with a determination of a significance Level of $P = 0.01$ for a t-test for a later application.

To select features which contain a significant difference, a SNR-like criterion was developed. It can be regarded as a similarity measure calculated by two partial averages. For data groups with partially disjoint features a sum-like and product SNR criterion was developed. The approach for these criteria was justified by a feature selection of artificial data. Here, a control group was used to study the performance yielding a confirmation of our approaches. Hence, the application of the presented methods for feature selection of TF transformed biomedical data seems to be well suited.

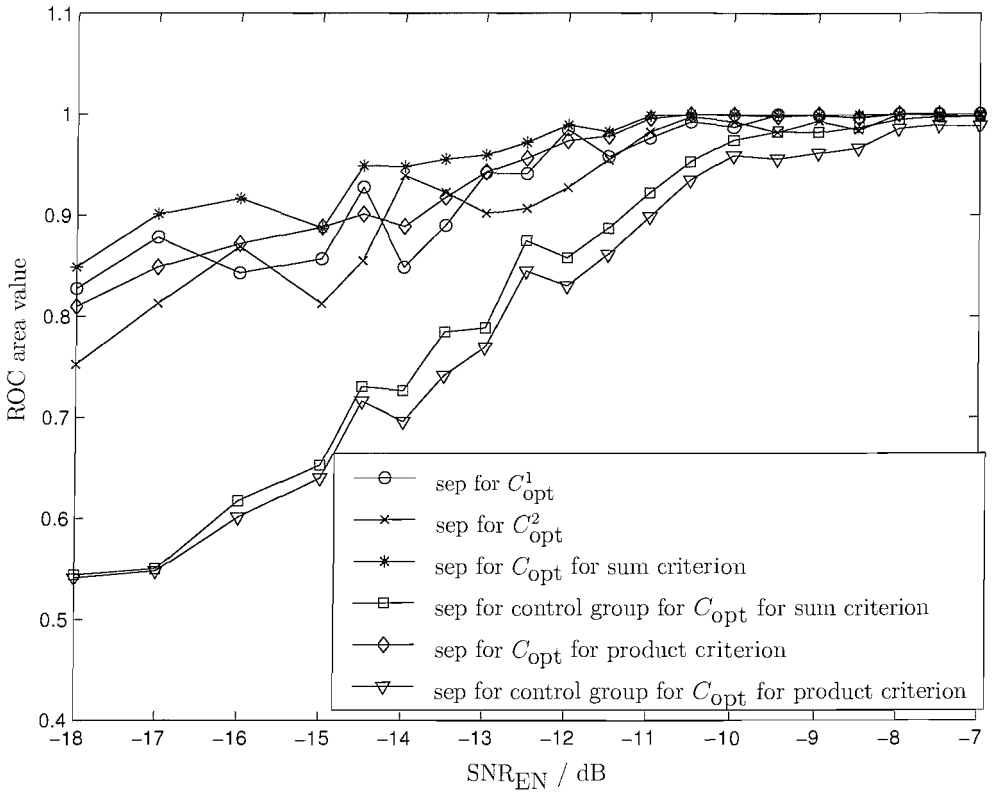

Figure 3.14: ROC area values based on the various criteria including control groups for neighbourhood search.

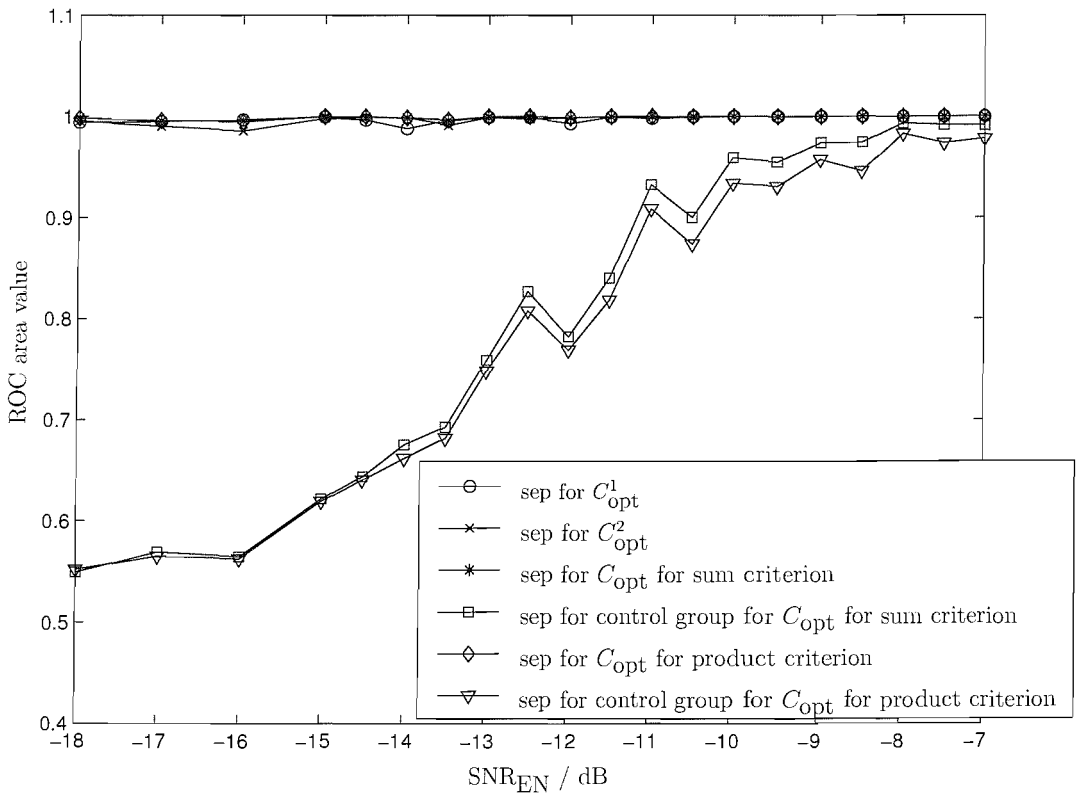

Figure 3.15: ROC area values based on the various criteria including control groups for exhaustive search.

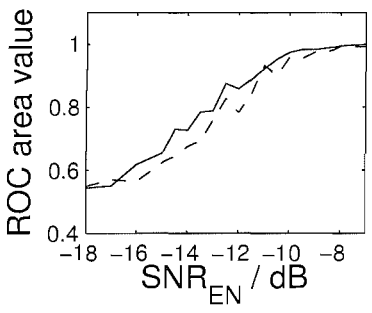

Figure 3.16: Comparison for control groups: ROC area values for sum-like combined criterion $\hat{\xi}^{+}$ for first and second order neighbourhood growth (solid) and for an exhaustive search (dashed).

Chapter 4

Classification: Support Vector Machines

Having described the feature selection method in the previous chapter, we continue with the explanation of the classification methods we explored. Considering Figure 1.1, we arrived at the last part on the right. As stated in the previous chapter, the ROC analysis provides an evaluation of a classification by linearly changing the threshold. In this chapter, we introduce a method that selects a threshold based on the selected features in a higher dimensional feature space aiming at an further improvement of the value given by the ROC curve. The obtained threshold can also be shifted in the higher dimensional space and hence, a ROC analysis of the classification can be conducted. Here, we are using support vector machines (SVM) as they are a relatively new classification method and their application to real-life data is still under investigation. For a detailed description, we refer to [9],[53],[54],[55],[56],[57],[58]. SVM are assumed to yield results competitive to neural networks (NN). As studies for the performance of NN exist for the type of data we analyse, we contribute to that research area by deploying SVM. The chapter is organised as follows: Firstly, a general introduction to learning theory is given by Section 4.1 to lay the foundation of learning machines such as SVM or NN. In Section 4.2, the SVM theory is introduced and their connection to learning theory is shown. These two sections are mainly based on the description of SVM in [54]. In Section 4.3, we introduce the adaptation of the theory to our application for diagnosis of potential diseases based on biomedical data. Section 4.4 gives a summary of the chapter.

4.1 Introduction to Learning Theory

The motivation for the development of learning theory can be regarded as the attempt to describe the problems with learning formally and based on that formulation, find solutions for those problems. The main issue regarding a learning machine is: How good does it generalise? As we will show later, the generalisation of a learning machine is closely connected to its capacity. Capacity is defined as the maximum number of points which can be correctly classified by a learning machine when the data points show the most unfavourable labels. We will give a more detailed definition of capacity later. If the capacity is too big, so-called overfitting can occur. If it is too small, underfitting can appear.

This can be explained by an example trying to identify a tree [54]. A botanist with a photographic memory may say, unless the object has the same number of leaves as one tree she can remember, it is not a tree, which is overfitting. The botanist's lazy brother may conclude, it is a tree, as long as it is green, which corresponds to underfitting.

4.1.1 Definitions: Training Data, Capacity Description

In the following, we define some terms for learning machines, starting with the training data which is a data set with corresponding labels:

$$(\mathbf{x}_1, y_1), \ldots, (\mathbf{x}_L, y_L), \tag{4.1}$$

where $\mathbf{x}_l \in \mathbb{R}^N$ are the training samples and y_l are the labels for the simple pattern recognition case $y \in \{\pm 1\}$. The distribution probability $P_p(\mathbf{x}, y)$ is unknown hereby. We can assume that the $\mathbf{x}_l$ describe a space χ in $\mathbb{R}$ which ought to be split into two parts based on the labels y_l. For a detection system, the $\mathbf{x}_l$ could hold the TF transformed data after a feature selection as explained in the previous chapters. Then, the data would be labelled into healthy or diseased for the detection of a disease e.g., where L represents the number of measurements taken.

A learning machine is a quantity of functions $\{f(\mathbf{p}_\alpha)\}$ with $f(\mathbf{p}_\alpha) : \chi \rightarrow \{\pm 1\}$, where $\mathbf{p}_\alpha$ is a vector holding parameters. A certain choice of those parameters contained by $\mathbf{p}_\alpha$ generates a trained machine, which separates the space χ into the classes -1 and $+1$. E.g. if $\{f(\mathbf{p}_\alpha)\}$ is a set of linear functions, $\mathbf{p}_\alpha$ would hold two parameters, the gradient and the offset from the origin. A certain selection of these parameters yields a trained learning machine. There are 2^L possibilities to label the L samples. If one function from the quantity $\{f(\mathbf{p}_\alpha)\}$ corresponds to one of those 2^L possibilities, the learning machine is said to shatter those L points. The Vapnik-Chervonenkis (VC)-dimension h_{VC} is defined as the maximal number of points that can be shattered by a learning machine. This means that the VC-dimension is used to describe the characteristic of f and is dependent on it, it is not dependent of a certain choice of the parameters $\mathbf{p}_\alpha$. The VC-dimension is a well-known measure for the capacity of a learning machine and will be explained in more detail by an example. To avoid confusion, h_{VC} is the only definition and measure for the capacity of a learning machine which will be used from now on.

Suppose the function set $\{f(\mathbf{p}_\alpha)\}$ consists of oriented straight lines to shatter two classes in the space $\mathbb{R}^2$. Those lines represent the separating hyperplanes. This learning machine has a VC-dimension of three, which is illustrated in Figure 4.1. In this example, the maximum number of points that can

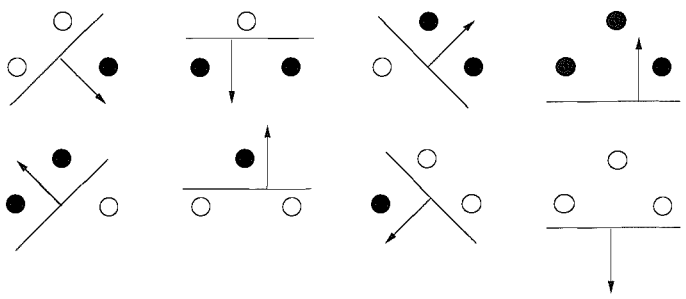

Figure 4.1: Oriented lines can shatter a maximum of three points in $\mathbb{R}^2$.

always be shattered by the learning machine is three except if the three points lie on a straight line. There are 8 possibilities how the three samples can be labelled and for each case there is a function $f(\mathbf{p}_\alpha)$. Four points cannot always be shattered by this function set. These exceptions are illustrated by Figure 4.2.

Figure 4.2: Oriented lines cannot shatter four points (left) or points that lie on a line (right) in $\mathbb{R}^2$.

Generally, a function set of oriented hyperplanes in $\mathbb{R}^N$ has a VC-dimension of $N+1$ [9].

Intuitively, one could assume that a learning machine with many parameters $\mathbf{p}_\alpha$ would possess a high VC-dimension, while a learning machine with few parameters would have a low VC-dimension. This is generally not the case. There exist learning machines with only one parameter and infinite VC-dimension. A learning machine with infinite VC-dimension is said to be able to shatter L points, no matter how large L is. For illustrative purposes, an example is given to explain these statements in more detail [54].

Example:

In the following, the training data vector $\mathbf{x}$ is a scalar denoted by x, the same applies to $\mathbf{p}_\alpha$.

Define the step function

$$\theta(x) = \begin{cases} 1, & x > 0 \\ -1, & x < 0. \end{cases}$$

Consider the function set with one parameter p_α:

$$f(x, p_\alpha) \equiv \theta(\sin(p_\alpha x)), \; x, p_\alpha \in \mathbb{R}.$$

For an arbitrary L, we choose x_l according to

$$x_l = 10^{-l}, \; l = 1, \ldots, L, \tag{4.2}$$

with arbitrary labels $y_1, y_2, \ldots, y_L, \; y_l \in \{\pm 1\}$.

Then, $f(p_\alpha)$ yields those labels for

$$p_\alpha = \pi(1 + \sum_{l=1}^{L} \frac{(1-y_l)10^l}{2}),$$

and therefore, $h_{VC}(\{f(p_\alpha)\}) \to \infty$.

However, although this learning machine has infinite VC-dimension, we can also find four points, that cannot be shattered, which are illustrated in Figure 4.3. The white points belong to one class, the black point to the other class. This shows, that the class points need to fulfil the condition defined by (4.2) for the learning machine to be able to shatter an unlimited amount of points. One could also say, that the infinite VC-dimension is due to an ill-chosen $f(x, p_\alpha)$. The purpose of this example is however to demonstrate that a learning machine with only one parameter can have infinite capacity. We continue with the definition of errors and their minimisation for a learning machine.

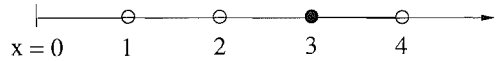

Figure 4.3: Four points that cannot be shattered by $\theta(\sin(\alpha x))$, although $h_{VC} \rightarrow \infty$.

4.1.2 Structural Risk Minimisation (SRM)

The actual risk is the expected error for the trained learning machine and given by:

$$R(\mathbf{p}_\alpha) = \int \frac{1}{2}|y - f(\mathbf{x}, \mathbf{p}_\alpha)| dP(\mathbf{x}, y) \tag{4.3}$$

which we aim to minimise. However, the class allocation probability $P_p(\mathbf{x}, y)$ is unknown. Therefore, we can only calculate the empirical risk which is the error for the training data and defined as:

$$R_{\text{emp}}(\mathbf{p}_\alpha) = \frac{1}{L}\sum_{l=1}^{L}\frac{1}{2}|y_l - f(\mathbf{x}_l, \mathbf{p}_\alpha)|. \tag{4.4}$$

The minimisation of R_{emp} does not automatically lead to a small actual risk. Figure 4.4 shows a learning machine for a minimised R_{emp} and different VC-dimensions emphasising the dependence of a minimised empirical risk on the capacity of the classifier. Therefore, when using empirical risk minimisation, we need to find a learning machine with an adequate capacity. Considering this issue, the statistical learning theory provides some bounds that limit the size of $R(\mathbf{p}_\alpha)$. One such bound is called risk bound and given by [9]:

$$R(\mathbf{p}_\alpha) \leq R_{\text{emp}}(\mathbf{p}_\alpha) + \phi_{VC}(h_{VC}, L, \delta_{VC}), \tag{4.5}$$

where $\phi_{VC}(h_{VC}, L, \delta_{VC})$ is called VC-confidence and is defined by

$$\phi_{VC}(h_{VC}, L, \delta_{VC}) = \sqrt{(\frac{1}{L}(h_{VC}(\ln(\frac{2L}{h_{VC}} + 1) + \ln\frac{4}{\delta_{VC}})} \; . \tag{4.6}$$

This bound is valid with a probability of δ_{VC}. It is independent of $P_p(\mathbf{x}, y)$ and can be effortlessly determined when h_{VC} is known. To obtain a well performing learning machine, we aim at minimising this risk bound. However, one has to keep the problems in mind that are connected with the theoretical risk bound $R(\mathbf{p}_\alpha) \leq R_{\text{emp}}(\mathbf{p}_\alpha) + \phi_{VC}(h_{VC}, l, \delta_{VC})$. For practical use, it often leads to non-trivial cases. Also, for a learning machine with infinite VC-dimension like the previous example, the bound is not even valid [9].

The VC-confidence depends on the function set and not on a certain choice of $\mathbf{p}_\alpha$, which will lead to difficulties if one aims at minimising it. This problem can be solved by introducing a structure for the function set meaning that we divide the entire function set $\{f(\mathbf{p}_\alpha)\}$ into nested subsets as shown in Figure 4.5.

The subsets are ordered by h_{VC}. Now, we just minimise the empirical risk for each subset. Then, the minimum of the risk bound is obtained by choosing the learning machine whose sum of empirical risk and VC-confidence is minimal. This approach is called Structural Risk Minimisation (SRM). Figure 4.6 shows the SRM of (4.5) graphically. It illustrates the general relation between capacity, training error and the bound on the test error: The larger the capacity namely the VC-dimension h_{VC}, the smaller the training error resulting in a learning machine too adapted to the training data and

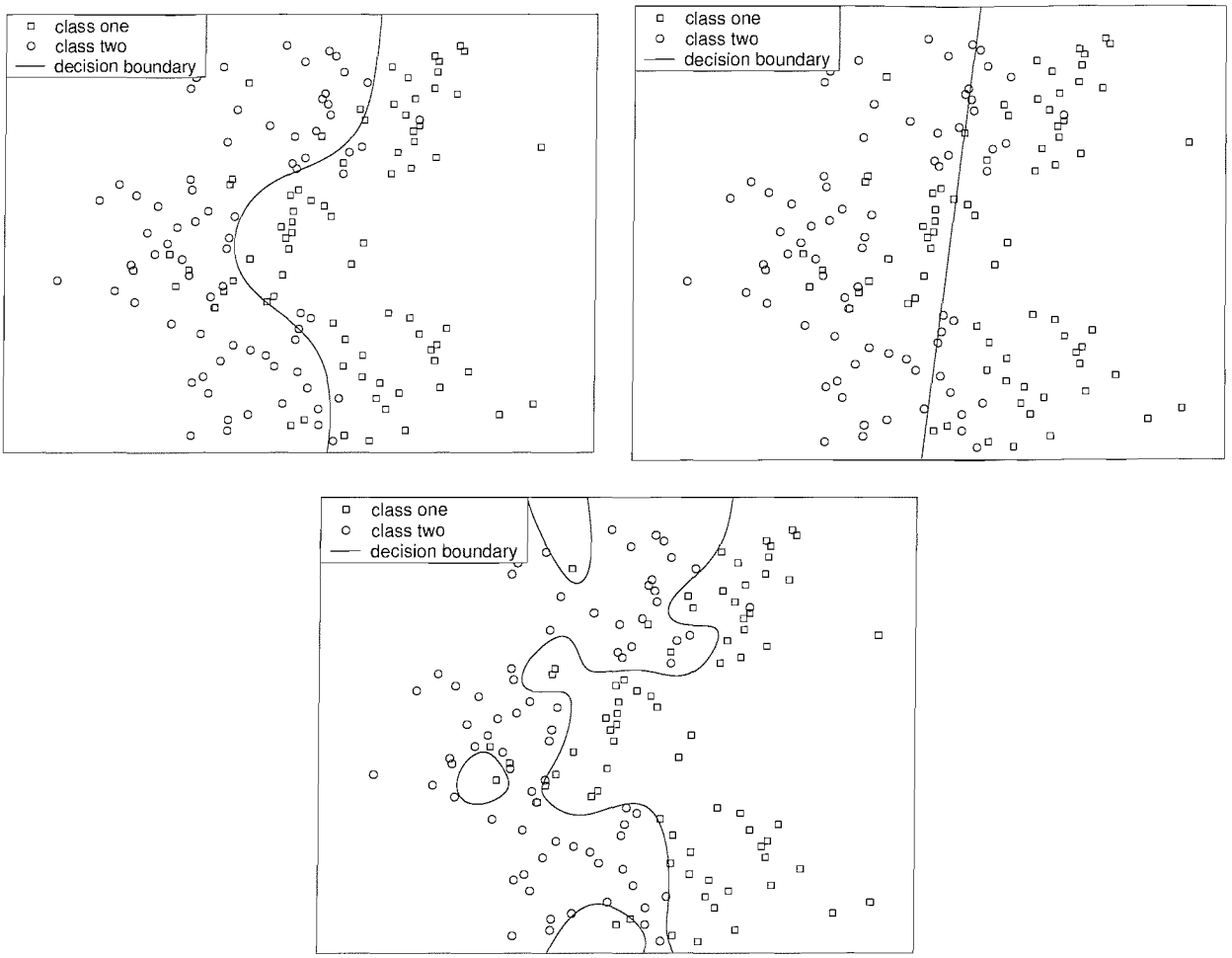

Figure 4.4: Learning machine with (top left) good classification, (top right) underfitting and (bottom) overfitting resulting from a minimised R_{emp} but different capacities (VC-dimensions).

therefore resulting in a high bound on the test error. This case is referred to as overfitting. For small capacities of a learning machine, already the training error shows relatively large values, which also results in a large value for the bound on the test error, and is known as underfitting. With SRM we aim at finding the optimum trade-off between both the training error and VC-confidence to minimise the error for the test data.

A learning machine with infinite VC-dimension can show a good performance, e.g. the k-nearest neighbour classifier with $k_k = 1$ [54]. This learning machine has infinite VC-dimension and zero empirical risk, as any number of points, which are labelled arbitrary, can be successfully learned by the classifier (provided that no two points of opposite class lie on top of each other). Therefore, the bound provides no information. But nearest neighbour classifier can still perform well. This example shows that infinite capacity does not guarantee poor performance.

Based on the above introduction to learning theory, we can now show the relation to support vector machines (SVM).

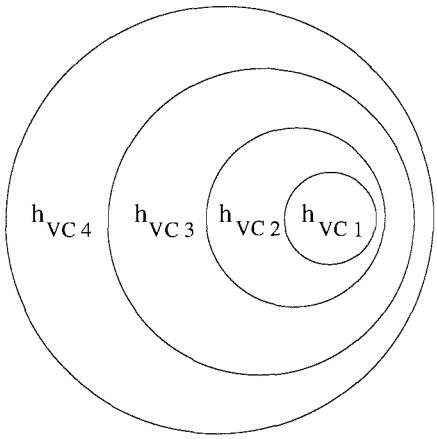

Figure 4.5: Nested subsets of function set $\{f(\mathbf{p}_\alpha)\}$, ordered by declining h_{VC}: $h_{VC\ 1} < h_{VC\ 2} < h_{VC\ 3} < h_{VC\ 4}$.

4.1.3 SRM, SVM and Neural Networks

SVM and neural networks (NN) can be regarded as instances of learning machines. Hereby, NN have a given structure by a number of layers, a number of nodes which can be seen as a type of non-linearity. The coefficients of the NN are optimised according to some cost function, e.g. the error on the training data which equals the empirical risk. SVM are applied based on a given empirical risk and the cost function according to which they are optimised, is the VC-confidence. They determine a separating hyperplane to which the distance for all data points for the two classes is the largest when all data points are accounted for. This statement is illustrated in Figure 4.7 for $\mathbb{R}^2$ for a linear learning machine. The two parameters in $\mathbf{p}_\alpha$ are the gradient and the offset from the origin. The offset from the origin is set to a constant value in the example. Then, the parameter which can be adjusted and optimised is the gradient. For the separating hyperplanes on the left, the sum of the distance for all data points to the corresponding separating hyperplanes is smaller than for the case on the right which represents the hyperplane with the largest distance when all data points are considered. Hence, the VC-confidence is minimised as the probability δ_{VC} for the bound $R(\mathbf{p}_\alpha)$ to be valid is maximised and enters (4.6) inversely. The data points needed for a definition of the separating hyperplane are called support vectors which can be seen as equivalent to the coefficients of a NN.

The SVM aim at finding the support vectors by applying a constraint on the size of the margin. Related to SRM, finding the maximal margin corresponds to minimise the VC-confidence for a certain empirical risk. Moreover, to find the minimal $R(\mathbf{p}_\alpha)$ also R_{emp} and h_{VC} can be varied. This will be explained in the next section where we continue with a more detailed description of the SVM theory.

4.2 SVM Theory

This section explains the SVM theory, starting with the simplest case of a linearly separable SVM, followed by the non-separable case and concluding with non-linear SVM classification.

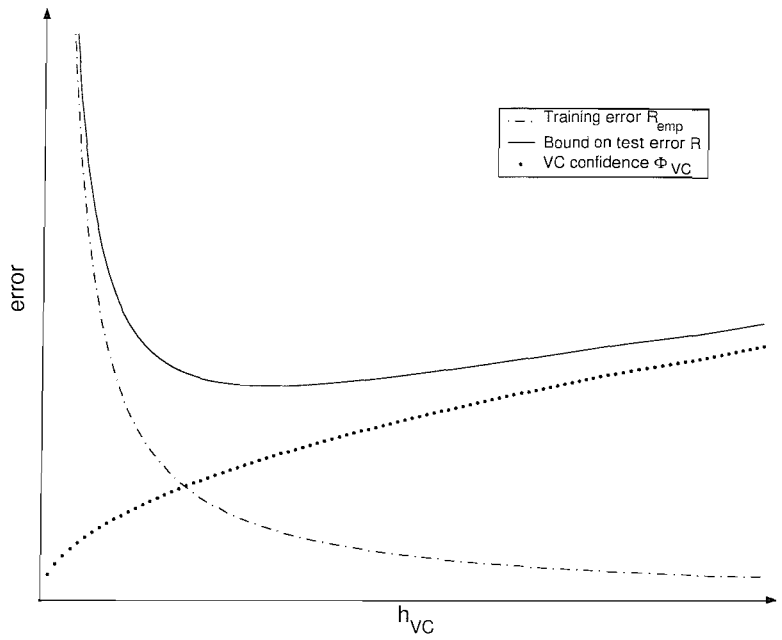

Figure 4.6: SRM for risk bound $R(\mathbf{p}_\alpha)$.

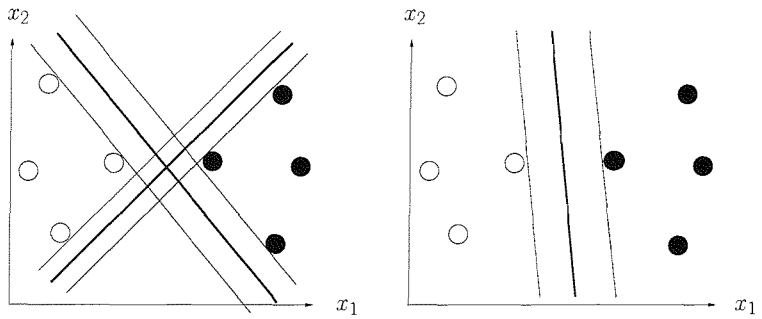

Figure 4.7: (Left) possible separating hyperplanes for class one ○ and class two ●; (right) separating hyperplane with maximal margin aimed at finding with SVM.

4.2.1 Linearly Separable Case

Let us assume that a two class learning problem is a given and the training data is defined according to (4.1) with the space $\chi = \mathbb{R}^N$. For the classification procedure it will also be assumed that supervised learning is conducted meaning that the allocation of the data to one of the classes is known and therefore, the classifier does not need to allocate single data to a class. For the data to be linearly separable, there exists a separating hyperplane for the data. This definition is illustrated in Figure 4.8 (left) for $\mathbb{R}^2$. For $\mathbb{R}^2$ the separating hyperplanes are straight lines.

To select one separating hyperplane among the possible ones, the hyperplane with the largest margin is sensible theoretically as well as illustratively. Figure 4.9 shows this hyperplane for the example in Figure 4.8 left. The hyperplane with the largest margin is defined by only a few data points which are called support vectors and are illustrated with dashed circles in Figure 4.9.

To describe the above statements mathematically, we define a hyperplane formally. A quantity of

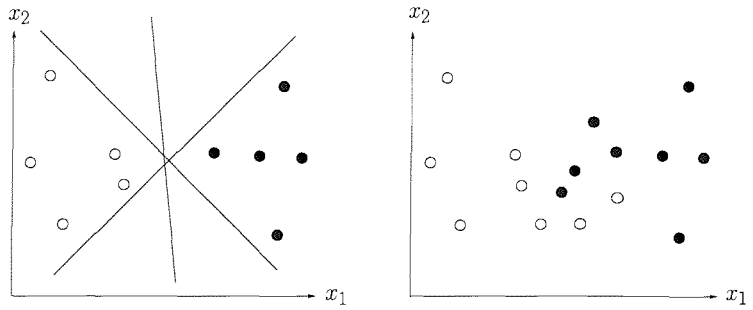

Figure 4.8: Data $\mathbf{x}$ for class one $\circ$ and class two $\bullet$: (Left) linearly separable data and (right) non-linearly separable data.

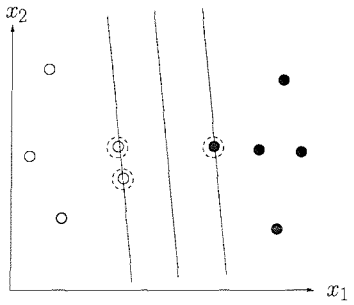

Figure 4.9: Separating hyperplane with maximal margin defining support vectors (dashed).

vectors $\mathbf{x}$ following

$$\mathbf{w}^H \cdot \mathbf{x} + b_s = 0, \ \mathbf{w}, \mathbf{x} \in \mathbb{R}^N, \ b_s \in \mathbb{R}, \tag{4.7}$$

with $\mathbf{w}^H$ being the Hermitian of an orthogonal vector $\mathbf{w}$ on the plane and b_s being the scaled distance from the origin; in more detail the term $\frac{|b_s|}{||\mathbf{w}||}$ defines the perpendicular distance from the origin with $||\mathbf{w}||$ being the Euclidean norm of $\mathbf{w}$. Figure 4.10 illustrates this definition for $\mathbb{R}^2$.

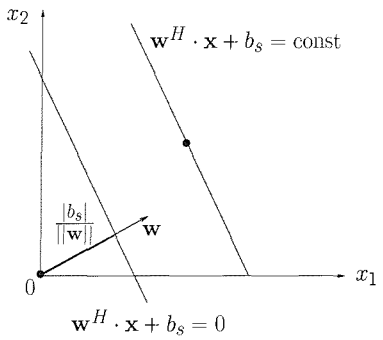

Figure 4.10: Definition of a hyperplane; setting the right side of (4.7) to a constant yields a parallel shifting of the hyperplane.

To separate data points by the hyperplane a linear classifier can be applied:

$$f(\mathbf{x}) = \text{sign}(\mathbf{w}^H \cdot \mathbf{x} + b_s) \tag{4.8}$$

According to (4.7), the hyperplane can be rescaled, e.g. $c \cdot \mathbf{w}^H \cdot \mathbf{x} + c \cdot b_s = 0$ with c being a constant. Based on this we can choose points closest to the hyperplane that satisfy $|(\mathbf{w}^H \cdot \mathbf{x}_l) + b_s| = 1$, whereby the margin equals $\gamma_m = \frac{2}{||\mathbf{w}||}$. Figure 4.11 shows these assumptions for $\mathbb{R}^2$. The points determined by

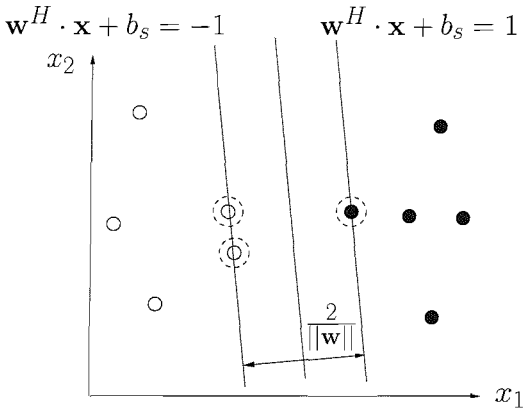

Figure 4.11: Normalised maximal margin for classification.

this approach are the support vectors which are circled in the figure. To find them the problem can be described as an optimisation:

$$\text{minimise } ||\mathbf{w}||, \text{ subject to } y_l(\mathbf{w}^H \cdot \mathbf{x}_l + b_s) \geq 1,$$

whereby the latter condition ensures a correct classification for the data labelled by y_l and an empty zone within the margin. According to [9] this optimisation problem can be described by a general Lagrange function according to:

$$L(\mathbf{w}, b_s, \alpha) = \frac{1}{2}||\mathbf{w}||^2 - \sum_{l=1}^{L} \alpha_l [y_l(\mathbf{w}^H \cdot \mathbf{x}_l + b_s) - 1], \tag{4.9}$$

with $\alpha_l \geq 0$. The task for finding the optimum is to minimise (4.9) with respect to $\mathbf{w}, b_s$ and to maximise it with respect to α_l [55]. Instead of solving this optimisation directly it is easier to solve the Lagrangian dual problem [9] according to

$$\alpha_{\text{opt}} = \arg\max_{\alpha} L'(\alpha) = \sum_{l=1}^{L} \alpha_l - \frac{1}{2} \sum_{l,i=1}^{L} y_l y_i \alpha_l \alpha_i (\mathbf{x}_l^H \cdot \mathbf{x}_i), \tag{4.10}$$

with $\alpha_l \geq 0$ and $\sum_{l=1}^{L} y_l \alpha_l = 0$. The solutions for this problem are obtained via so called Karush-Kuhn-Tucker conditions (KKT) [9].

Then, the solution α_{opt} of the dual problem provides the wanted hyperplane:

$$\mathbf{w} = \sum_{l=1}^{L} \alpha_{\text{opt},l} y_l \mathbf{x}_l. \tag{4.11}$$

There is also a solution for b_s which is implicitly determined by the KKT [53].

According to (4.9), for all data vectors that are located outside the margin, $\alpha_l = 0$ and hence, this can be denoted as a sparse representation of the solution. If $\alpha_l \neq 0$, the respective data vector $\mathbf{x}_l$ is a support vector. Moreover, the hyperplane is unique and represents the global optimum of the Lagrangian dual maximisation.

Now, the determined support vectors $\mathbf{x}_l \; \forall \; \alpha \neq 0$ can be used for classification of a data vector $\mathbf{x}$ by substituting $\mathbf{w}$ in (4.8) resulting in the following decision function:

$$\begin{aligned} f(\mathbf{x}) &= \text{sign}(\mathbf{w}^H \cdot \mathbf{x} + b_s) \\ &= \text{sign}\left[\textstyle\sum_{l:\forall \alpha_l \neq 0} \alpha_l y_l (\mathbf{x}_l^H \cdot \mathbf{x}) + b_s\right]. \end{aligned} \tag{4.12}$$

Having shown how the use of support vectors can reduce the complexity of the decision function for classification, we proceed form the linearly separable case to the non-separable case next.

4.2.2 Non-linearly Separable Case

For data that cannot be separated linearly as shown in Figure 4.8 on the right, we introduce a slack variable ξ that can be regarded as penalty for errors. Figure 4.12 illustrates ξ.

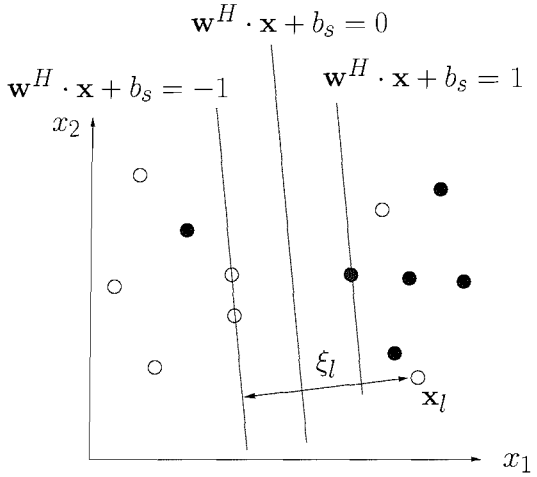

Figure 4.12: Illustration of slack variable ξ for data classification which is not linearly separable.

The formal description of the non-linearly separable case becomes:

$$\text{minimise} \quad ||\mathbf{w}||^2 + C \sum_{l=1}^{L} \xi_l, \tag{4.13}$$

with $y_l(\mathbf{w}^H \cdot \mathbf{x}_l + b_s) \geq 1 - \xi_l$ and $\xi_l \geq 0$. The parameter C can be chosen by the user, a larger C corresponds to assigning a higher penalty to errors. The dual formulation similar to (4.10) equals:

$$\alpha_{\text{opt}} = \arg\max_{\alpha} L'(\alpha) = \sum_{l=1}^{L} \alpha_l - \frac{1}{2} \sum_{l,i=1}^{L} y_l y_i \alpha_l \alpha_i (\mathbf{x}_l^H \cdot \mathbf{x}_i), \tag{4.14}$$

with $C \geq \alpha_l \geq 0$ and $\sum_{l=1}^{L} y_l \alpha_l = 0$. Therefore, the only difference to the linearly separable case is the condition $C \geq \alpha_l \geq 0$. The solution for this optimisation accords with the linearly separable case resulting again in a unique global optimum. When illustratively compared to the linearly separable case, the margin is not empty for the non-linear separable case and an increasing C leads to a smaller margin. Also, the identified support vectors are not only located on the margin but can also lie within the margin or can be wrongly classified data points which leads to qualitatively more support vectors than for the separable case. We call support vectors that lie on the margin bounded support vectors and the rest unbounded support vectors.

Next, non-linear SVM are introduced.

4.2.3 Non-linear SVM

The basic concept of non-linear SVM is the transformation of the data in the input space into a higher dimensional feature space where a linear SVM classification is conducted as explained above. Let Φ

be the transformation function, then $\Phi(\mathbf{x}) : \mathbb{R}^N \mapsto F$ maps the input space to the feature space F. Figure 4.13 gives an example where the data in the input space can be regarded as an exclusive OR (XOR) wiring.

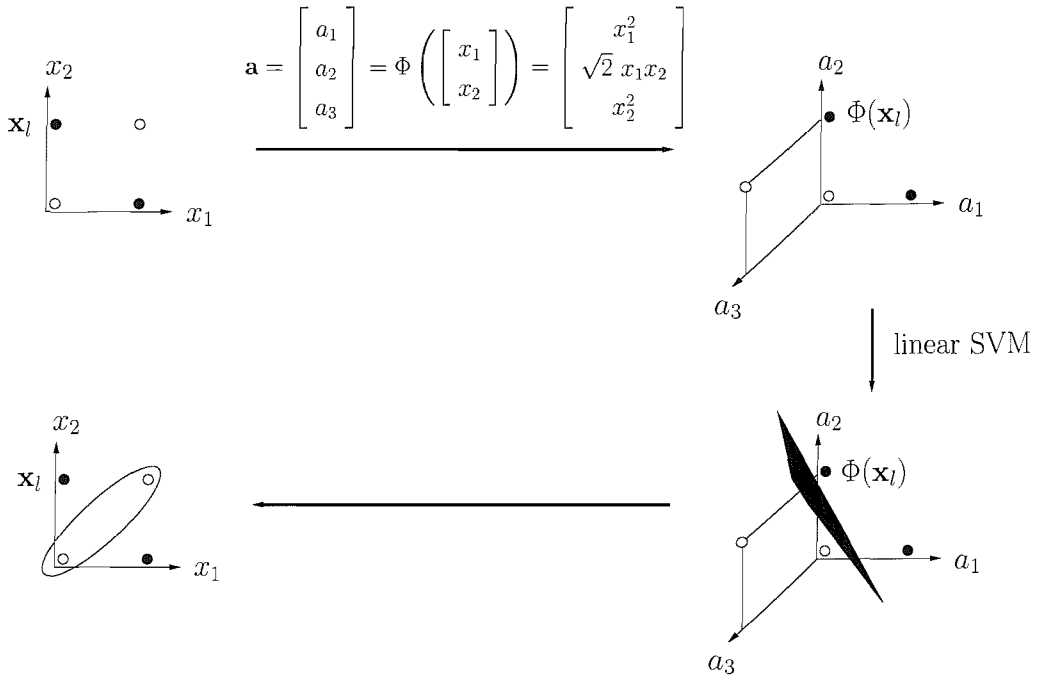

Figure 4.13: Mapping of a XOR wiring from $\mathbb{R}^2$ to $\mathbb{R}^3$ where a linear separating hyperplane can be found.

The training data is mapped by Φ to a higher dimensional space, where a linear separation can be conducted which corresponds to a non-linear separation in the input space. However, the calculation of the hyperplane with maximal margin in F can be quite complex. Investigation of (4.12) and (4.14) reveals the most important property of support vector machines: For training and classification only dot products need to be calculated. For the non-linear case, the products $(\mathbf{x}_l^H \cdot \mathbf{x}_i)$ in (4.14) would be substituted by $(\Phi(\mathbf{x}_l)^H \cdot \Phi(\mathbf{x}_i))$. However, these products can be substituted by efficient kernels according to $K_{SVM}(\mathbf{x}_l, \mathbf{x}_i) = \Phi(\mathbf{x}_l)^H \cdot \Phi(\mathbf{x}_i)$. This substitution does not require the knowledge of Φ or F, and can be directly computed in the input space. This so called kernelisation is applicable to numerous linear algorithms.

The kernel functions can be regarded as a measure of similarity or a general scalar product. The most common kernels are the following:

- linear: $K_{SVM}(\mathbf{x}_l, \mathbf{x}_i) = \mathbf{x}_l^H \cdot \mathbf{x}_i$;
- polynomial: $K_{SVM}(\mathbf{x}_l, \mathbf{x}_i) = (\mathbf{x}_l^H \cdot \mathbf{x}_i + c)^{d_p}$ with d_p being the order and c a constant;
- sigmoid: $K_{SVM}(\mathbf{x}_l, \mathbf{x}_i) = \tanh(\kappa(\mathbf{x}_l^H \cdot \mathbf{x}_i) + \Theta)$ with κ being the gain and Θ the offset;
- Gaussian: $K_{SVM}(\mathbf{x}_l, \mathbf{x}_i) = \exp(-\gamma_R ||\mathbf{x}_l - \mathbf{x}_i||^2)$ with γ_R specifying the radius of the radial basis function (RBF) kernel.

These kernels fulfil Mercer's condition [9] saying that any kernel which is positive definite can be expanded in its Eigenfunctions which is a requirement for the kernelisation.

E.g. in Figure 4.13, a polynomial kernel of order two is shown according to

$$\begin{aligned} K_{SVM}(\mathbf{x}_l, \mathbf{x}_i) &= (\mathbf{x}_l^H \cdot \mathbf{x}_i)^2 = \left(\begin{bmatrix} x_{l_1} \\ x_{l_2} \end{bmatrix}^H \cdot \begin{bmatrix} x_{i_1} \\ x_{i_2} \end{bmatrix} \right)^2 = \\ &= \left(\begin{bmatrix} x_{l_1}^2 \\ \sqrt{2}\, x_{l_1} x_{l_2} \\ x_{l_2}^2 \end{bmatrix}^H \cdot \begin{bmatrix} x_{i_1}^2 \\ \sqrt{2}\, x_{i_1} x_{i_2} \\ x_{i_2}^2 \end{bmatrix} \right) = \\ &= (\Phi(\mathbf{x}_l)^H \cdot \Phi(\mathbf{x}_i)) \end{aligned}$$

with $\Phi(\mathbf{x}) = \begin{bmatrix} x_1^2 \\ \sqrt{2}\, x_1 x_2 \\ x_2^2 \end{bmatrix}$.

Note that neither the mapping $\Phi(\mathbf{x})$ nor the space F are unique for a given kernel. The above mapping is just one example for a polynomial kernel of order two [54].

Now we can give the final formulation for the SVM theory. A support vector learning machine classifier is obtained by a training procedure according to:

$$\alpha_{\text{opt}} = \arg\max_{\alpha} L'(\alpha) = \sum_{l=1}^{L} \alpha_l - \frac{1}{2} \sum_{l,i=1}^{L} y_l y_i \alpha_l \alpha_i K_{SVM}(\mathbf{x}_l, \mathbf{x}_i), \tag{4.15}$$

under the constraints $C \geq \alpha_l \geq 0$ and $\sum_{l=1}^{L} y_l \alpha_l = 0$. To classify a data vector $\mathbf{x}$ with the so obtained learning machine the following equation is evaluated yielding a binary classification:

$$f(\mathbf{x}) = \text{sign}\left(\sum_{l:\forall \alpha_l \neq 0} \alpha_l y_l K_{SVM}(\mathbf{x}_l, \mathbf{x}) + b_s \right). \tag{4.16}$$

Now, we are able to state the relation between the parameters for SVM and the learning theory described in the previous chapter. The empirical risk R_{emp} is connected to the parameter C. The VC-dimension which describes the capacity of the learning machine is related to kernel parameters, e.g. for Gaussian RBF, via γ_R the capacity is determined; for a polynomial kernel, the order d_p restricts the capacity. RBF show an infinite VC-dimension h_{VC} as an infinite amount of points that can be classified correctly, as long as γ_R is large enough. By decreasing γ_R an overfitted learning machine applying RBF kernels can lead to a good classification as illustrated in Figure 4.14. It shows that changing γ_R changes the capacity related to h_{VC} of the learning machine, or in more detail, γ_R limits the maximal capacity of the learning machine, over which a constraint is applied during the training.

Therefore, by restricting the maximal capacity h_{VC} of a learning machine by kernel parameters, the classifier found by maximising the margin for a given empirical risk minimises the VC-confidence. Hence, related to SRM, R_{emp} can be changed by C, along with h_{VC} by kernel parameters and for each adjustment of these parameters, the VC-confidence is minimised. Hence, SRM can be practically conducted by minimising the number of support vectors when trying to find the optimum for C and h_{VC}.

The general architecture of SVM is shown in Figure 4.15. The input vector $\mathbf{x}$ and the support vectors $\mathbf{x}_l$ are non-linearly mapped into a higher dimensional space via Φ. In this space, dot products

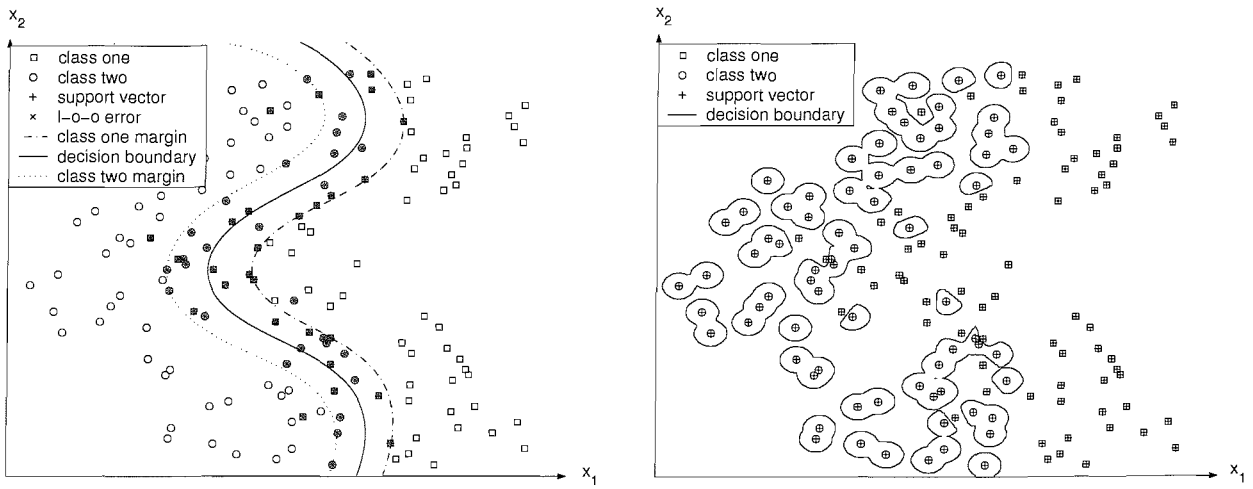

Figure 4.14: RBF learning machine with well adjusted capacity (left) and too large capacity (right) where all vectors are identified as support vectors.

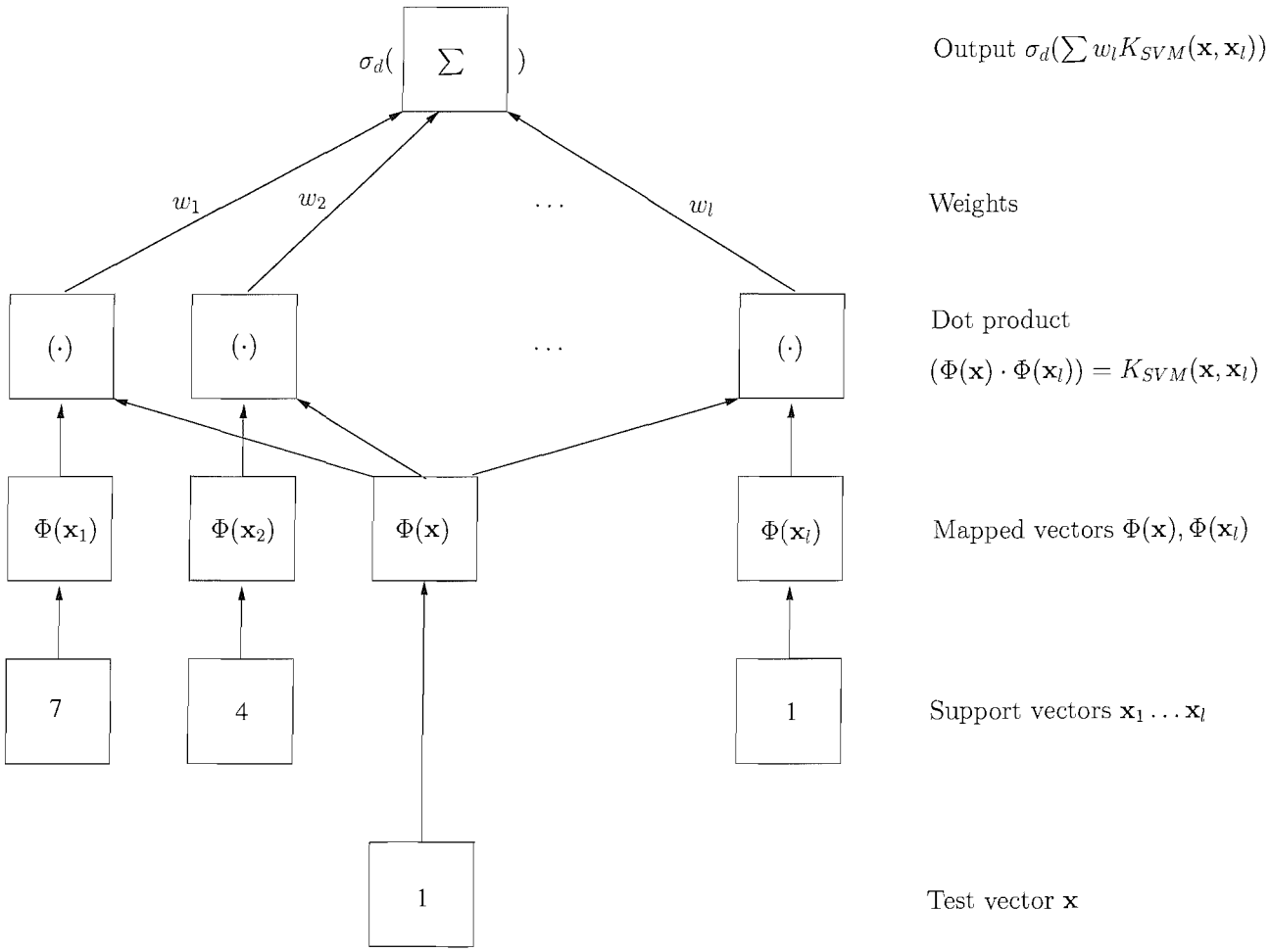

Figure 4.15: Overview: Architecture of SVM.

are calculated. By the use of kernels, these steps are combined, without any knowledge of the feature space F. The results are linearly combined weights w_l, resulting from solving a convex quadratic optimisation problem with $w_l = y_l \cdot \alpha_l$. The linear combination represents the input into a decision function $\sigma_d(x)$ with $\sigma_d(x) = \text{sign}(x + b_s)$.

One popular algorithm to conduct the training and find the support vectors $\mathbf{x}_l$ and weights w_l is called sequential minimal optimisation (SMO) [53]. There are quite a few implementations available on the internet from different research institutes, we have chosen to use [59] for our application.

Having introduced the SVM and shown their connection to learning theory, one major question remains to be answered: How do we conduct the SRM with SVM or in other words: How do we find the optimal parameters C and kernel parameters like γ_R or d_p with respect to the chosen kernel? The answer to this question is the calculation of a leave-one-out (l-o-o) estimation of the error rate as follows [9]: From the training samples, remove the first example. Train the SVM on the remaining samples. Then test the removed example. If the example is classified incorrectly, it is said to produce a leave-one-out error. A classifier with a low l-o-o error is supposed to possess a good generalisation avoiding overfitting, see Figure 4.14. Figure 4.14 shows the following statement illustratively: Only vectors identified as support vectors obtained by training on the complete data can influence the l-o-o error [57]. Therefore, the pure number of support vectors is an indication of a l-o-o error prediction: The fewer support vectors, the lower the maximal possible l-o-o error. As stated before, the support vectors consist of the points that lie on the margin, inside the margin or are falsely classified vectors.

In [57], a slightly narrower approach to estimate the maximum l-o-o error is shown avoiding training the SVM more than once, which is also used for our studies contained in this thesis. In general the estimation of the l-o-o error in [57] tends to overestimate the true error rate. However it provides a very computationally effective estimation and a more restricted solution than a pure count of the support vectors. Note that, in Figure 4.14 all support vectors are estimated to produce a l-o-o error.

Now, we have all the necessary tools to apply SVM: We are able to evaluate a certain choice of empirical risk by the parameter C and the limitation of the capacity by kernel parameters like γ_R or d_p with an estimation of a l-o-o error.

We continue with a description of how we apply SVM to real-life data.

4.3 SVM for Diagnosis

Firstly, we will give an outline how to apply support vector machines to a certain application. As the SVM conduct a sign decision between two classes we also address the following questions which arise for real-life data: How do we apply SVM e.g. to a diagnostic test? How do we adapt SVM to get a predetermined significance result for one parameter of a diagnostic test? Furthermore, can we determine soft decisions meaning that we reject some data vectors classifying them as neutral to achieve a certain significance result? Moreover, how do we deal with more than two classes?

4.3.1 Specific Application

When applying SVM to data sets, a suitable data representation must be chosen first. This means the data has to be scaled to avoid that data in greater numeric ranges dominate those in smaller numeric ranges. The authors of [60] suggest a linear scaling of the input data to range of $[-1, +1]$ which we adopt according to:

$$\mathbf{x}_{l,\text{ scaled}} = \frac{\mathbf{x}_l}{\arg \max_i ||\mathbf{x}_i||_\infty} \tag{4.17}$$

for $i \in C_T$ with C_T being the set of training data. We pick out the largest overall modulus $||\mathbf{x}_i||_\infty$ for the entire training data set by "arg max" and divide all training vectors by it. This is also known as the infinity norm which selects the largest row sum for a matrix and hence, when dealing with vectors, represents the maximal element of a vector.

After scaling, a kernel for the learning machine must be chosen. Standard kernels are listed in the previous section, with the polynomial and Gaussian being the most popular and promising ones [55]. Also, one could design and implement a kernel for a specific application, however this is beyond our scope here.

To determine the optimal values for the training error and the restriction of the capacity, the authors of [60] suggest a grid search varying the kernel parameters against C on a digital grid. The result for training is evaluated upon the l-o-o error estimation. The classifier showing the lowest l-o-o error is chosen as the best learning machine. The so-obtained classifier can then be applied to the test data.

At this point, let us give an example how we apply SVM also explaining SRM in more detail. Assuming we use a polynomial kernel, the number of bounded support vectors as defined in 4.2.2 is increased by increasing d_p and hence, also h_{VC} is increased as illustrated in Figure 4.5. Note that the kernel parameters are not part of the parameters $\mathbf{p}_\alpha$ of a learning machine. The parameters in $\mathbf{p}_\alpha$ that are optimised are the gradient and the offset of the separating hyperplane in the higher dimensional feature space. Now, R_{emp} is changed by C and adjusts the number of unbounded support vectors. Therefore, the grid search for C and d_p corresponds to the SRM shown in Figure 4.6, finding the minimum for $R(\mathbf{p}_\alpha)$ which again is found by minimising the VC-confidence. Also, as stated before, just selecting the values for C and d_p which show the minimum number of support vectors represents SRM, as they show the minimum for bounded and unbounded support vectors. However, applying a l-o-o error estimation is more restrictive.

4.3.2 Application of SVM for Diagnosis

The SVM defined in the previous section conducts a hard decision meaning that the sign function (4.16) allocates a data vector to one class. This decision is based on a boundary, found by an minimisation of the l-o-o error for generalisation purposes. When we think of a two class classification problem solved by SVM, we can connect that to a diagnostic test as it is defined in Section 3.1. Let us assume we conduct a diagnostic test similar to Section 3.1 with SVM: Class one consists of positive subjects, class two comprises negative subjects. The decision boundary found by the SVM can be evaluated by the rates shown in Table 3.1. E.g. Figure 4.16 shows the relation for a SVM classification and TP,

FP, TN and FN rates. Based on the decision boundary as shown in the figure, the TP equals 88%,

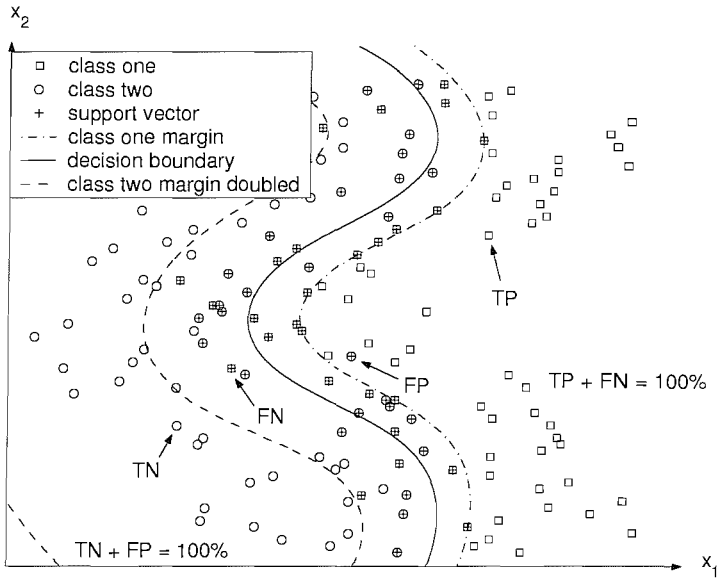

Figure 4.16: Connection of a SVM classification with a diagnostic test.

the TN rate 86.7%, the FP is 13.3% and the FN rate is 12%.

For such tests, one may select a certain threshold for one rate. E.g. for test on cancer, the falsely negatively classified rate should not exceed a certain value, say 2% even if that meant a higher falsely positively classified rate as it is better to wrongly classify people as having cancer than as falsely classify people as being healthy. We have developed a method how to set a threshold for one diagnostic test evaluation parameter, which we will show next.

4.3.2.1 Decision Boundary Shifting

To set a threshold for one of the rates TP, TN, FP or FN, we can just shift the decision boundary in the space F. In Figure 4.16 the class two margin is shown in the input space for a doubled margin for class two in the space F. If it is used as a decision boundary, there will be only one FN data vector compared to nine for the illustrated decision boundary. However this has to be traded off against the TN rate, which will decrease if the doubled two class margin is used as decision boundary. The shifting of the decision boundary is conducted in the higher dimensional space F of the SVM. According to Section 4.2.1 the margin for SVM classification can be rescaled to an interval of $[-1; 1]$, where 0 determines the decision boundary, see Figure 4.11. Therefore, by using a value of e.g. -2, the nonlinear decision boundary in the input space is changed equalling a parallel shifting of the hyperplane in the space F. Basically, we adjust the parallel shifting of the hyperplane in the space F by the FN rate parameter in the input space.

Clearly, shifting the hyperplane in F shifts a threshold over the data which can be analysed using a ROC. To illustrate the behaviour of the test rates for our example in Figure 4.16, Figure 4.17 illustrates the ROC curves for changing the decision boundary from -5 to 5 for the separability (TN) and FN rates.

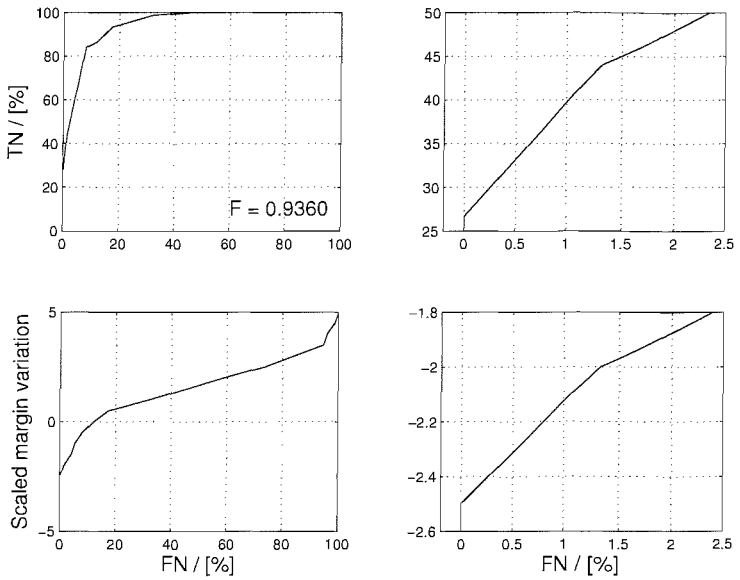

Figure 4.17: (Top) ROC curves: (left) TN vs FN, (right) zoomed in; (bottom left) run of the shift parameter for the decision boundary in the space F vs FN, (bottom right) zoomed in.

The right of the figure illustrates the zoomed in area of the FN rate from $0 \ldots 2.5\%$. We see that for achieving a FN rate of below 2% (e.g. classifying only 2% or fewer positive people as negative), the margin needs to be -2 which is illustrated in Figure 4.16 and the TN rate decreases from $\approx 86.7\%$ to $\approx 44\%$ which is the trade off for the low TN rate.

Based on the above explanations, we continue with the introduction of neutral decisions.

4.3.2.2 Neutral Class

For a diagnosis based on real-life data one can think of a decision, where the classifier gives a neutral result meaning that certain data points cannot be allocated as they are too close at the decision boundary. Hence, the introduction of a neutral class seems reasonable. Combined with the idea of setting a certain threshold for one diagnosis evaluation parameter, a neutral class can be defined by shifting the separating hyperplane in the higher dimensional space, starting from the decision boundary until a certain threshold is reached. This idea explained with the above example would lead to a neutral class illustrated in Figure 4.18.

Compared with the decision boundary shifting, the FP test evaluation parameter does not suffer so much from the determination of the threshold because part of it is classified as neutral. Table 4.1 shows the test evaluation parameters for the example illustrated in Figure 4.18 where the neutral class stretches from the original decision boundary found by the SVM applying a minimisation of the l-o-o error to twice the original margin of class two. Also, the table shows the respective values for the basic SVM classification and the decision boundary shifting for comparison reasons.

It is indicated that the introduction of a neutral class seems to be a good compromise when setting a threshold for one evaluation parameter for a diagnosis.

We continue this section with multi-class SVM.

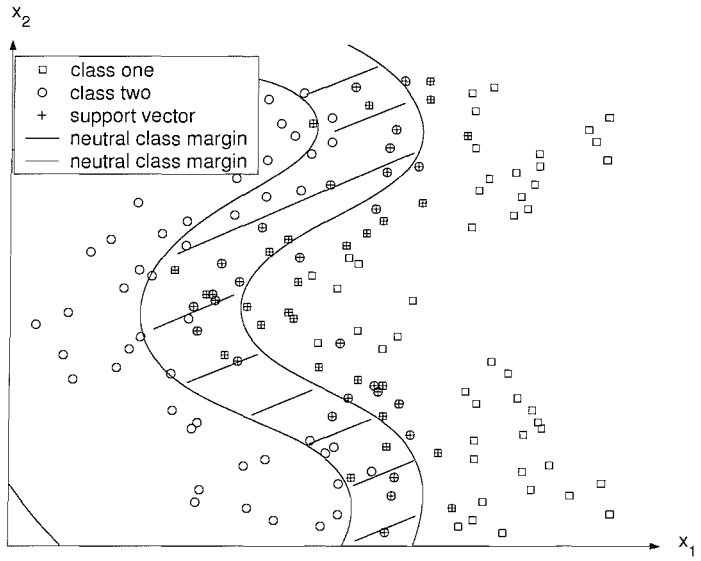

Figure 4.18: Neutral class yielding a FN rate of 2% obtained by shifting the separating hyperplane starting from the decision boundary to twice the original margin of class two.

	original SVM classification		shifted decision boundary to achieve FN of 2%		neutral class spanning from original decision boundary to threshold achieving FN of 2%	
	class one (positive group)	class two (negative group)	class one (positive group)	class two (negative group)	class one (positive group)	class two (negative group)
correctly classified (TP | TN)	88%	86.7%	98.7%	44%	88%	44.7%
neutrally classified	–	–	–	–	10.7%	42%
wrongly classified (FN | FP)	12%	13.3%	1.3%	56%	1.3%	13.3%

Table 4.1: Results comparison for our example showing all test evaluation parameters.

4.3.3 Multi-Class SVM

We have shown how to apply SVM for diagnosis including the rejection of a decision by a neutral class. Another important issue is the multi-class problem [61],[62],[9]. For a medical diagnosis application we can think of a test that aims at detecting different severities of an illness, e.g. different categories of hearing loss. As multi-class SVM is an area of strong interest in the research community, we give a very brief introduction and proceed straightforwardly to the method we have chosen. The solution for a multi-class SVM classification problem can be split into two groups:

- a combination of several two class SVM, or
- a integration of all data points, so-called all together methods.

For the first case, three popular methods which are applied for pattern recognition e.g. in [61], are [62]:

- one-versus-rest (1-v-r), where each class is classified against the rest;
- one-versus-one (1-v-1), where all classes are classified against each other;
- directed acyclic graph (DAG).

For the second case, a general method is suggested by [9] based on a 1-v-r method. However, the classification is achieved by the solution of only one equation in contrast to the 1-v-r method, where each class is tested against the rest.

In [62] a decision DAG for multi-class SVM is introduced. It is based on an 1-v-1 classification where the training is conducted for all possible combinations of the classes. Based on a trained SVM classifier for each possible class combination, a binary acyclic graph is used for testing. Figure 4.19 shows a decision DAGSVM for four classes.

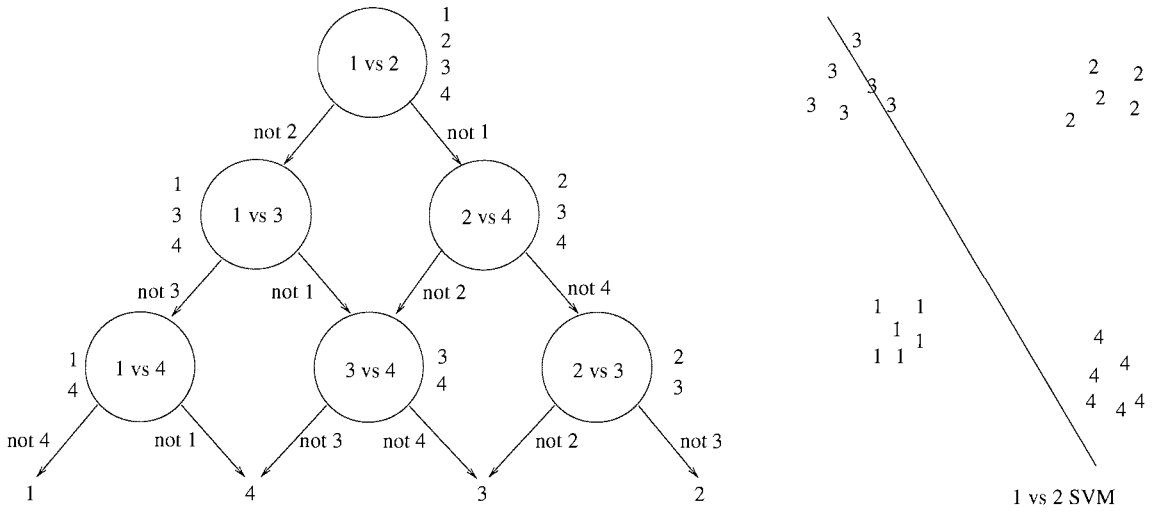

Figure 4.19: (Left) DAGSVM for four classes and (right) sample 1-v-1 SVM for training.

According to [62] the DAGSVM shows the same results as a 1-v-r or 1-v-1 multi-class SVM classification, however a much faster computation, as fewer tests are needed to arrive at a decision. Also,

all together methods do no outperform DAGSVM in terms of computational time [9]. These are the reasons why we have chosen DAGSVM for our application as a fast computational method based on a binary 1-v-1 SVM classification seems sensible.

Next, we introduce neutral decisions for DAGSVM.

4.3.4 Neutral Class for Multi-Class SVM

We have shown how to set a threshold for a neutral class and how to deal with multi-classes. Now, we combine the two methods by introducing a neutral decision for each node of the DAGSVM. To illustrate the extreme case, we define the neutral class as large as possible meaning we aim at a FN and FP rate of zero for the training data. Figure 4.20 explains the idea for three classes for the training.

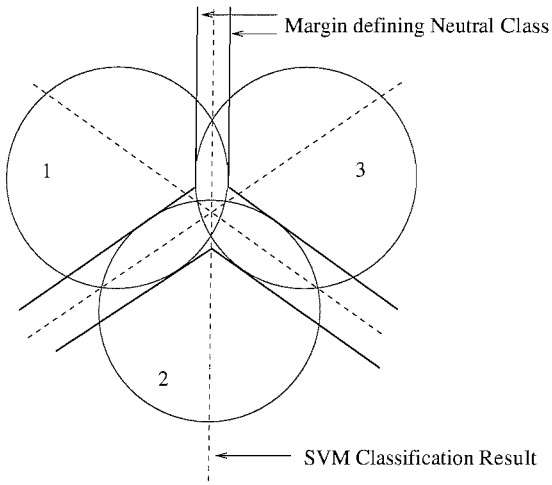

Figure 4.20: Definition of neutral classes for mulit-class SVM classification.

For a later DAGSVM the dotted lines in Figure 4.20 determine the classification result for the training data. For the illustrated neutral class, a neutral decision is yielded for all data points that are classified incorrectly for the training data. However, as explained above, the cost for this can be a severe decrease in the TN and TP rates.

The DAGSVM classification for testing is altered according to Figure 4.21 for the example for three classes. For the application of the introduced neutral DAGSVM a smaller threshold than the maximal possible can be used to determine the neutral class. Moreover, the DAGSVM decision tree can be altered according to the specific application to be analysed. In the figure, if there is a neutral outcome at the first node, the node in the middle described by " 2 vs 3 & 1 vs 2" will conduct an AND, meaning that a decision for class 2 will only occur if "2 vs 3" yields 2 AND "1 vs 2" yields 2. All other outcomes will be allocated to the neutral class. Furthermore, the decisions in the middle include a neutral class meaning that the result for "2 vs 3" and "1 vs 2" in the middle node can also be neutral.

Recapitulating it can be said that the introduced approach seems promising to be a good compromise yielding neutral decisions for multi-class SVM.

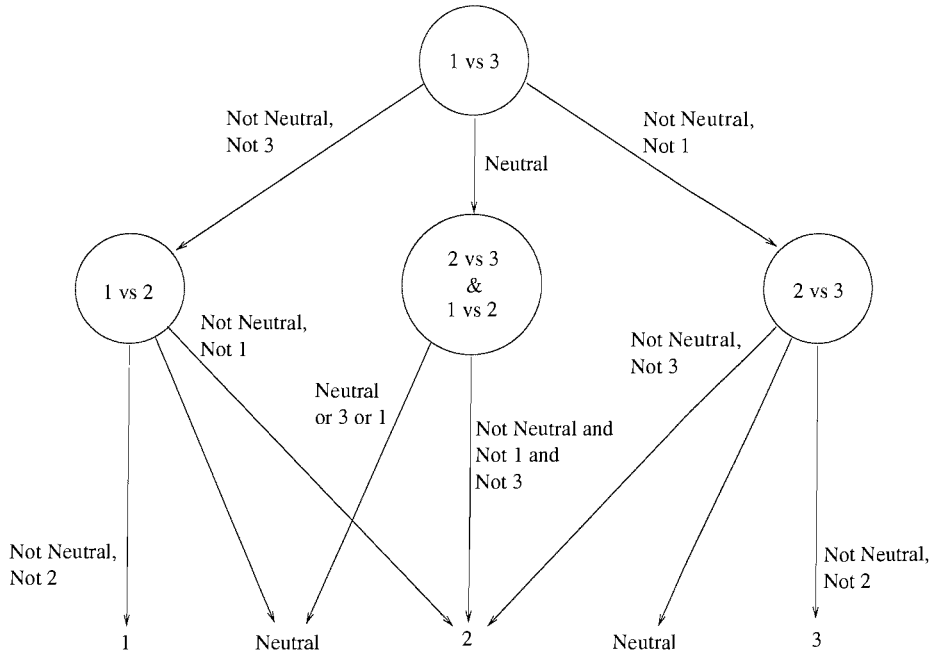

Figure 4.21: DAGSVM with neutral decisions for three classes.

4.4 Summary

In this section, we have introduced SVM and shown their connection to learning theory by a mathematical formulation of the training procedure, capacity limitation, classification and generalisation. We have also discussed SVM for the application for diagnosis introducing neutral decisions for multi-class classification. We conclude this section by giving an overview of the drawbacks and advantages of SVM from a practical point of view.

One drawback is the high computational effort for standard techniques for learning and classifying. Also, there is no completed multi-class expansion, this issue is reduced to binary two class problems. Moreover, a proposition for classification reliability is missing. Finally, there exists no common criteria for selecting a certain kernel.

The advantages of SVM are their illustrative functionality and that they are empirically explainable. Furthermore, they are based on theory whereby the classifier is found by determining a guaranteed global optimum which also allows an evaluation of generalisation. By fast implementations (SMO), the computational effort can be reduced. In addition, SVM are easily applicable, there are only few parameters too choose and only little a-priori-knowledge is required. Finally, the solution describes a sparse representation.

Chapter 5

TF Transform Based Feature Selection for Panic Disorder

In medical facilities it is a common issue to judge the responses of subjects to stimuli in order to determine a potential physiological or psychological illness [5]. In cases where the response can be measured as an electrical signal, the signal evaluation used to be primarily based on an expert's decision regarding the waveforms of averaged signals. Such waveforms and often parameters derived from these as presented by standard clinical measurement devices were treated as additional information only. Recently, however, automated evaluation methods based on signal processing approaches have more and more frequently enhanced or even replaced the expert's judgement [8],[4],[5]. As a result, many different propositions were made concerning statistical signal evaluation in an effort to enhance or perhaps even replace the decision of a human expert.

In the following two chapters, we contribute to evaluate the separation of biomedical data to describe potential illnesses by showing the application of the TF and feature selection methods introduced in Chapter 2 and 3. In this chapter, we analyse panic disorder based on event-related brain potentials (ERP) that are part of the electroencephalogram (EEG). ERP can be used to differentiate between the responses to neutral or panic disorder triggering stimuli when presented to anxiety patients. Moreover, we employ TF revealing transforms and statistical tests to identify a small number of significant parameterising coefficients that permit us to perform — as well as quantify — this differentiation.

The chapter is organised as follows. Section 5.1 gives an introduction to EEG and panic disorder ERP. Based on some properties of ERP and a parameterisation by WP and GF analysis, Section 5.2 contains the results and discussion of the feature selection by statistical tests as shown in Section 3.3 as only data from one patient was available. As the introduced approach of analysing panic disorder is a novelty, Section 5.3 justifies our approach by comparing our results with other methods like a simple time domain average, a KLT transform or applying the statistical tests to unparameterised data. Finally, a summary is given in Section 5.4.

5.1 Introduction

5.1.1 Electroencephalogram (EEG) and Its Sources

The electroencephalogram (EEG) [1] is a relatively low-frequency (30 Hz) recording of a voltage signal taken from the scalp which reflects electrical activity of the brain. The first recordings were made by Hans Berger in 1929 [1]. The amplitude varies from as large as 150 μV peak-to-peak to 50 μV peak-to-peak which is more common. EEG can be used to diagnose brain abnormalities and also to localise brain volumes responsible for certain mental activities. Also, by applying visual or auditory stimuli, the response EEG can be used in psychology for diagnosis and psycho-physiological research [8].

In order to understand how the electrical activity of the brain can be disrupted, it is first necessary to have a basic idea of how this signalling works [1]. The brain is essentially a mass of cells, or neurons, which are capable of transmitting chemical signals to one another and propagating electrical signals internally. This information can be transmitted because neurons are each equipped with several special features that make them different from other cells.

All neurons have a cell body and several extensions from this cell body called dendrites and axons. A typical neuron usually has many dendrites and only one very long axon. The dendrites contain receptors embedded in the membrane, which are proteins that respond to chemical signals. The axon has a terminal at its end that releases the chemical messengers, or neurotransmitters, to the next neuron.

One of the most important ideas to consider about signal transmission in neurons is that there is an actual physical space between one neuron's axon and the next neuron's dendrites. This space is called the synapse and is the site of communication between cells. The neuron releasing the neurotransmitter is called the pre-synaptic neuron and the one accepting the neurotransmitter is called the post-synaptic neuron. The pre-synaptic neuron will release its neurotransmitter into the synapse. This chemical messenger will diffuse across the synapse and interact with specific receptors on the post-synaptic membrane.

The interaction between the neurotransmitter and receptor will cause changes in the post-synaptic neuron which will cause the electrical potential (or signal) mentioned above to be generated. This electrical signal will be driven down the length of the axon until it reaches the terminal. When the signal hits this part of the axon it will cause the terminal to release neurotransmitter into the next synapse. These neurotransmitters will cross the synapse and interact with the receptors on the next neuron, thus continuing the process.

All neurons in the brain generally function in this manner, but it is also important to realize that there are as many as about 100 billion neurons in the brain. The question arises how these electrical signals of the brain can be used to describe our abilities as human beings. When the brain is damaged by a tumour or stroke, for example, the electrical signals can't get through to their appropriate connections and the function of the brain is impaired which can be observed by measurements via an EEG. Therefore, in the next subsection, we show how the EEG is recorded.

5.1.2 EEG Recording

The EEG is measured from electrodes which are placed in special positions on the scalp. These positions are identified by the recordist who measures the head using the international 10/20 system [1]. This relies on taking voltage measurements between certain fixed points on the scalp. The EEG electrodes are placed on the scalp at 10 and 20 percent of certain measured distances. If a measurement was made around a skull, (e.g. a hat size measurement) a circumference distance would be determined. This measurement would be around 55 cm for an average-sized person. Therefore, 10% of this measurement or 5.5 cm would be used to determine precise locations around the skull. An anatomic landmark is needed to know where to start with the measurements. These landmarks are the ear canals, the bridge of the nose (nasion) and the inion at the very back of the skull.

The 10/20 system is based generally on the relationship between the location of an electrode and the underlying area of cerebral cortex, although exact relationships can only be determined with further confirmation of exactly where the various parts of the brain are. There are slight variations amongst individuals in brain shape and relationship to the skull landmarks. Each point, e.g. C_z, C_3, F_{p1}, F_7, etc. represents a standard place for a recording electrode. Each site has a letter to identify the lobe and a number or another letter to identify the hemisphere location. The letters F, T, C, P, and O stand for frontal, temporal, central, parietal and occipital lobes. Note that there is no "central lobe", but this is just used for identification purposes. Even numbers (2,4,6,8) refer to the right hemisphere and odd numbers (1,3,5,7) refer to the left hemisphere. The z refers to an electrode placed on the mid-line. The reference or ground electrode is usually placed on the ear lobe, A_1 or A_2 [63]. Figure 5.1 illustrates a sample location of the electrodes. Also note that the smaller the number, the closer

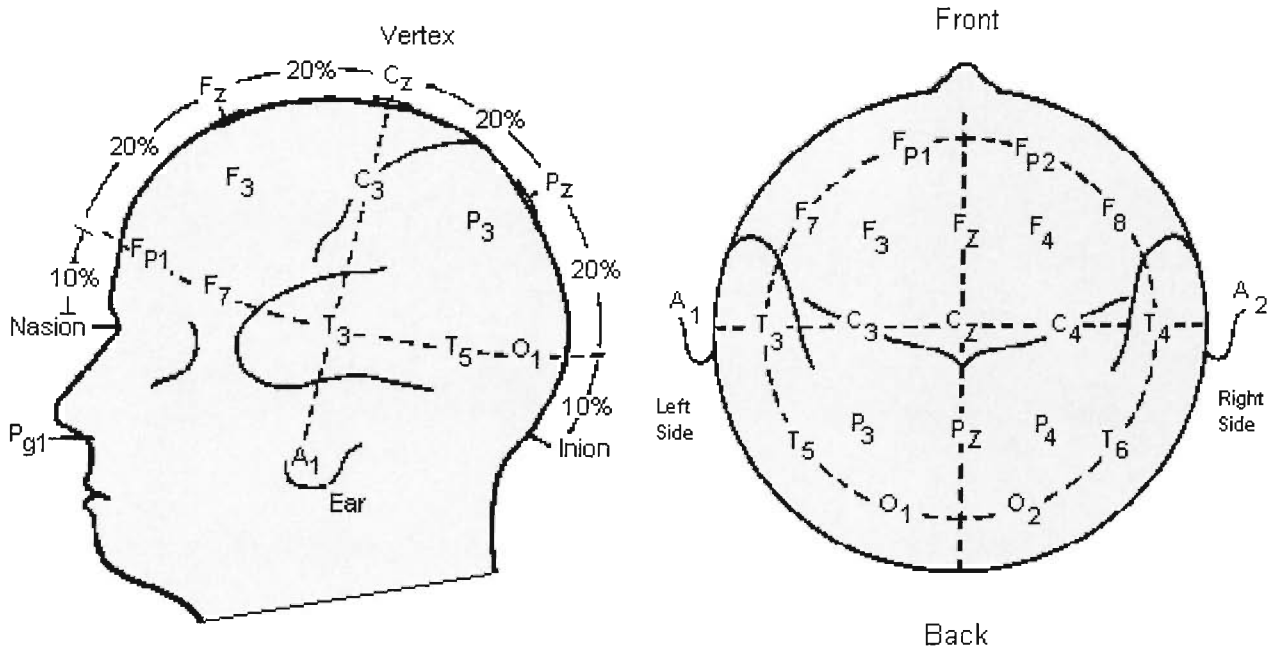

Figure 5.1: Sample location of the electrodes for EEG measurement according to the 10/20 system (with permission of BrainMaster Technologies, Inc., 24490 Broadway Avenue, Oakwood Village, Ohio 44146, USA for academic use).

the position is to the mid-line.

Many more placements can be defined for research purposes. Some researchers use 64, 128, or even 256 electrode placements [1]. Figure 5.2 shows a sample subject with EEG electrodes placed on her skull.

Figure 5.2: EEG measurement (with permission of BrainMaster Technologies, Inc., 24490 Broadway Avenue, Oakwood Village, Ohio 44146, USA for academic use).

We continue with a description of the various wave types found in an EEG.

5.1.3 EEG Classification

Spontaneous EEG activity can be broken down into 4 distinct frequency bands [1]:

- Beta activity $\geq$ 13 Hz,
- alpha activity 8 Hz – 13 Hz,
- theta activity 4 Hz – 7 Hz and
- delta activity $\leq$ 4 Hz.

Beta activity is a normal activity present when the eyes are open or closed. It tends to be seen in the channels recorded from the centre or front of the head. Some drugs will increase the amount of beta activity in the EEG.

Alpha activity is also a normal activity present in waking adults. It is mainly seen in the channels recorded from the back of the head. It is fairly symmetrical and has an amplitude of 40 μV to 100 μV. It is only seen when the eyes are closed and should disappear or reduce in amplitude when the eyes are open.

Theta activity can be classed as both a normal and abnormal activity depending on the age and state of the subject. In adults it is normal if the subject is drowsy. However it can also indicate brain dysfunction if it is seen in a subject who is alert and awake. In younger subjects, theta activity may be the main activity seen in channels recorded from the back and central areas of the head.

Delta activity is only normal in an adult subject if they are in a moderate to deep sleep. If it is spontaneously seen at any other time it would indicate brain dysfunction.

Another activity that can be measured by the EEG are event related brain potentials [1] which are described next.

5.1.4 Event Related Brain Potentials (ERP)

ERP are the evoked transient EEG response to sensory stimulation either visual or auditory or tactile (transient pressure applied periodically to body parts). They can be used for diagnosis in psychology. Transient electrical activity is evoked by the applied transient stimuli; firstly from the sensory nerve input nuclei, then from the brain-stem and after this from the sensory cortex. ERP are generally small, often around 1 μV peak. Therefore, ERP possess the same order of magnitude as amplifier noise and noise from muscles picked up by the EEG electrodes [1]. Furthermore, ERP can be smaller than other unrelated EEG activity measured by the respective electrode.

Synchronous signal averaging is a common method to extract ERP transients out of the additive noise [1]. Here, we will use TF transformations to extract the features of the ERP. In the next subsection, the panic disorder ERP are described.

5.1.5 Panic Disorder ERP studies

Individuals with panic disorder are characterised by an abnormal fear of certain anxiety connected sensations such as palpitation, breathlessness, or dizziness [64]. The research into this disorder has led to studies investigating its symptoms by means of appropriate stimulation and measurement of the subsequent ERP [65, 66]. In this context, visual stimulation has been performed with words causing panic disorder, whereby the EEG can be recorded showing event related potentials. Previous studies have resulted in revealing a low frequent transient waveform appearing approximately 300 ms after stimulus onset as a distinctive characteristic which is referred to as P300.

Signal averaging followed by analysis of variances (ANOVA) is commonly used in detecting the P300 and studying panic disorder based on the response ERP [14],[3]. Since the P300 has a transient behaviour, the application of TF analysis appears well suited, as it takes both spectral and temporal information into account [67]. In the next section we aim to investigate various transforms — such as wavelet, wavelet packet, and Gabor transforms — with respect to their suitability for revealing the TF characteristics of the transient P300. We further optimise these transforms such that the distinction between panic disorder and normal responses is concentrated in only few transform coefficients, to which we apply a statistical test for the feature selection as only little data is available.

5.2 Feature Selection for Panic Disorder

In this section, we introduce the experimental conditions under which panic disorder data was obtained. Furthermore, TF transforms will be reviewed, which can parameterise the elicited event related potentials. Finally, test results for the described methods are presented and discussed including the application of the statistical tests according to Section 3.3.

5.2.1 Description of Data

For the measurement of panic disorder ERP, an anxiety patient was presented with fear-inducing or neutral words briefly at the perception threshold of panic disorder. The patient's perception threshold

for correctly identifying 50% of the words was determined with neutral words not used in the experiment. Based on the assumption that he/she will recognise a greater number of anxiety words given at his/her perception threshold than neutral words, the hypothesis examined is the expectation that his EEG exhibits an enhanced P300 wave for presented anxiety words [14].

The EEG was measured at the vertex electrode (C_z, see Figure 5.1) synchronously to the stimulus, whereby the recordings were started 100 ms before the onset of the visual word stimulus. The data analysed in this study contains 24 neutral word presentations and 24 anxiety word presentations to one panic patient. Figure 5.3 shows the average over the stimulus-synchronous EEG in reaction to the 24 words presented for each word category. There is a visible difference in the two averages with a stronger P300 and more positive EEG until approximately $t = 700$ ms in the panic disorder related data.

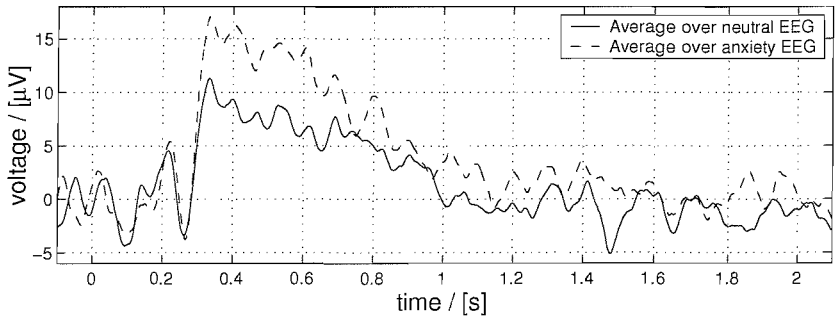

Figure 5.3: Average over 24 EEG segments showing responses to anxiety related and neutral stimuli at the perception threshold.

To emphasise the use of TF transforms on which the feature selection will be based, Figure 5.4 shows the averages over the neutral and anxiety words separately including the 68% confidence bands, meaning the single standard deviation added and subtracted for each average. As the bands are relatively wide, the application of statistical tests and TF transforms seem to be a sensible approach also to locate the right frequency ranges for the P300.

5.2.2 Parameterising Transforms

To parameterise the ERP in Figure 5.3, TF transforms lend themselves to account for the transient nature of the waveforms. To capture the impulsive rise of the P300, TF transforms with a good time resolution are required. The DWT generally yields a good frequency resolution and poor time resolution at low frequencies, yielding a too coarse time segmentation in the frequency range of interest which is due to the dyadic TF tiling, see Figure 2.6. The only adjustment that can be made is the appropriate selection of the mother wavelet. Therefore, instead we consider the WP transform, whose level of decomposition can be adapted to fit the nature of the data, as well as the Gabor transform, which yields a uniform tiling of the TF plane and hence can provide a desired resolution in a specific TF segment.

Based on the implementation described in Chapter 2, the WP uses Mallat's wavelet [67], whereby the decomposition level of the transformation is adapted to minimise the entropy of the average ERP curves in Figure 5.3. The Gabor transform is based on an oversampled filter bank with 32 channels

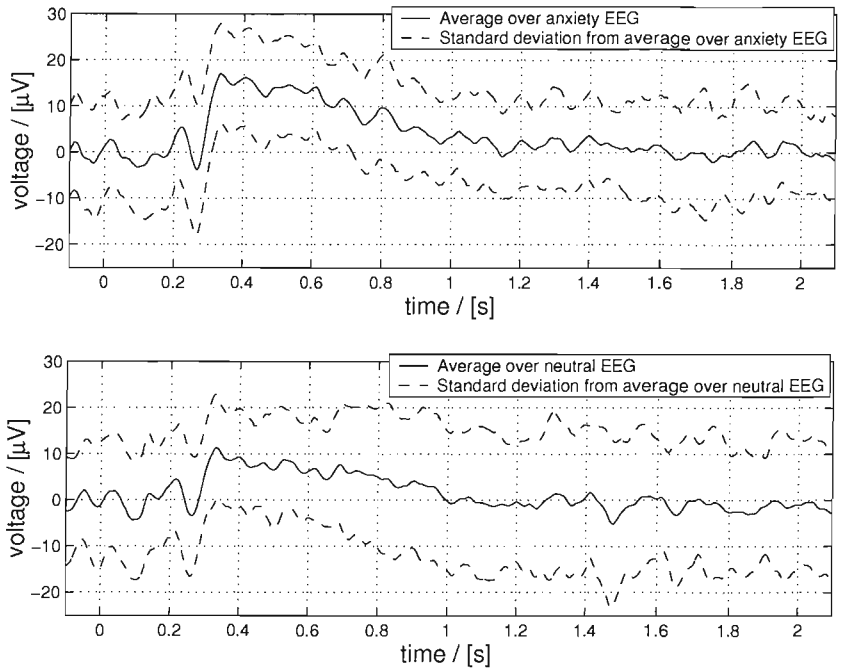

Figure 5.4: 68% confidence bands for (top) anxiety EEG and (bottom) neutral EEG.

constructed according to Section 2.5. The resulting approximate distribution of the coefficient energies in the TF plane is visualised in Figure 5.5.

The application of the transform methods leads to a parameterisation of the ERP data whereby the features of the ERP are expressed in as few coefficients as possible. Within these ERP-parameterising coefficients, we isolate those that represent a significant difference between the two data sets by using the statistical tests described in Section 3.3 as only 24 measurements for each data category are available.

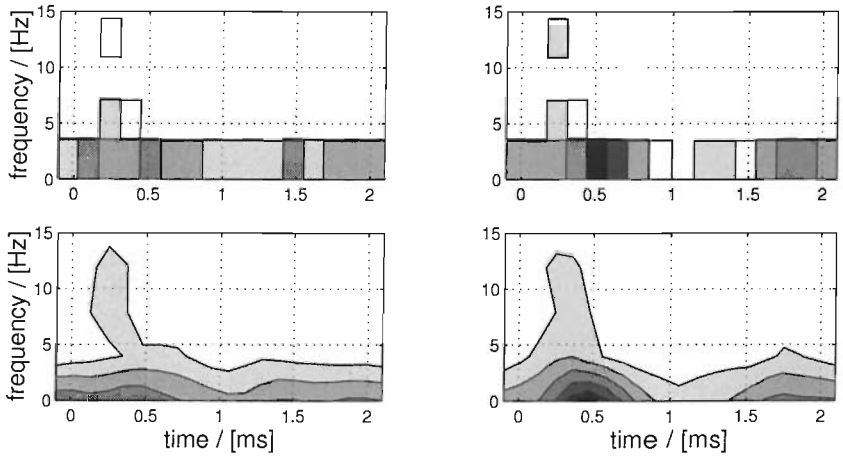

Figure 5.5: Average coefficient energy for (left) neutral words and (right) panic order related words using (top) WP and (bottom) Gabor transforms.

5.2.3 Results and Discussion

As discussed in 5.2.2 and 3.3, we have different transform methods and a procedure to identify significant coefficients that enables us to separate between presented neutral and anxiety words. In the following, we will discuss the used transforms and present the results for separability which we obtained for the data described in 5.2.1.

Transform Adjustment

The optimal decomposition structure for the WP is found over minimising the entropy as described in Section 2.4. The decomposition depth was limited to have at least 16 coefficients in one decomposition level as further decomposition would lead to a too coarse time segmentation. In terms of the Gabor transform, various filters were tested and it was found that using a prototype filter with length of $N_F = 224$, a frequency segmentation of $J = 32$ uniform scales and a time segmentation of $D = 14$ for the oversampling shows the best results for the separability which is determined by the area under the ROC curve. Table 5.1 shows the area under the ROC curve results for various prototype filters to explain how the described filter was selected.

Channel number J	decimation ratio D	length of prototype filter N_F	separability (area under the ROC curve) of "best" coefficient
64	56	448	0.7135
64	28	448	0.6944
64	14	448	0.7066
32	28	224	no coefficient found
32	14	224	0.7388
32	7	224	0.7108
16	14	112	0.7033
16	7	112	0.7050
8	7	56	0.7052

Table 5.1: Results for tested prototype filters for GF transform to separate presented panic and neutral words.

Identified Coefficients and Difference Comparison

The coefficients to which the statistical tests are applied were preselected whereby only coefficients are considered which contain 85% of the total energy according to (3.3), which is reasonable, as it reduces the probability of identifying coefficients that contain noise only. The value of 85% results from not considering coefficients that are located above 15 Hz in the TF plane, see Figure 5.5.

Figure 5.6 shows the resulting coefficients when performing a difference evaluation on the parameterised data, by statistical tests. We see that two coefficients (black and grey) for both transforms are identified. They cover approximately the area of the P300 slow wave as it is expected in 5.2.1. They are all identified via a t-test according to a prior F-test whereby the threshold for the significance level for the F-test was $P = 0.05$, and for the t-test, it was set to $P = 0.01$ according to Section 3.3.3.

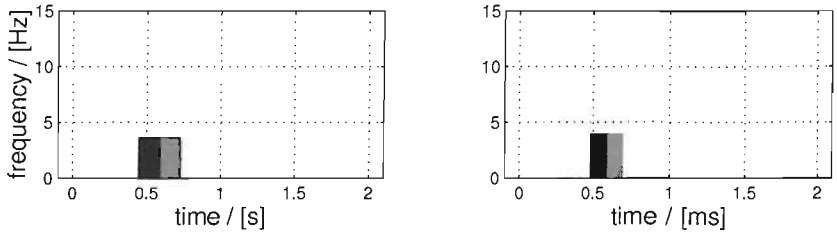

Figure 5.6: Resulting coefficients for (left) WP and (right) Gabor transforms.

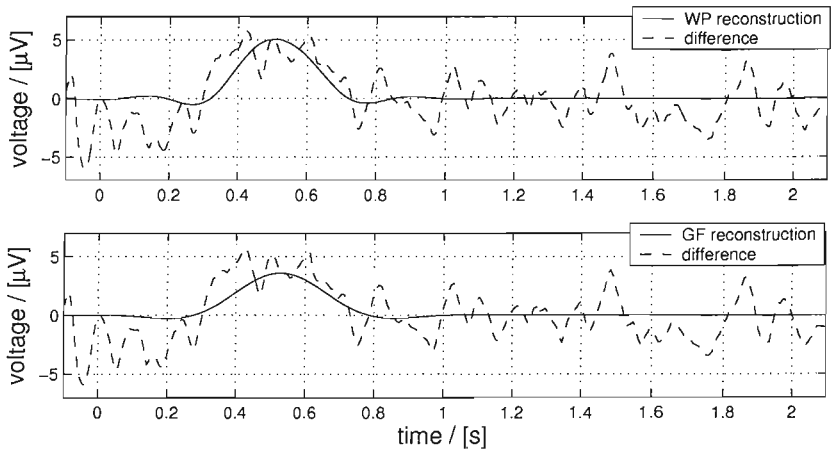

Figure 5.7: Difference of raw neutral and anxiety EEG data compared with its parameterisation by the two identified coefficients for (top) WP and (bottom) Gabor transforms.

Figure 5.7 shows the difference of the averages of the raw neutral and anxiety EEG compared with its parameterisation by the identified coefficients for the two investigated transforms. It can be observed that the two identified coefficients parameterise the P300 area very well for both transforms.

To evaluate the separability of the identified coefficients, Table 5.2 indicates the ROC curve analysis for these coefficients. All coefficients obtained show an equal or greater value than 0.72 which can be

	Transform	
Coefficient	WP	GF
black	0.73	0.73
grey	0.72	0.72

Table 5.2: Area under ROC curve for the identified coefficients.

regarded as a reasonably good discrimination. Recapitulating, it can be said that with both transforms an adequate separation of data of both categories, namely presented neutral and anxiety words, can be achieved. To support this statement, we show the separability for a simple time domain average, other transforms like the DFT and KLT and a comparison with unparameterised data for further justification.

5.3 Justification of TF Transform Based Feature Selection for Panic Disorder

5.3.1 Simple Time Domain Average

One might ask, observing the difference between the time average for the neutral and anxiety data in Figure 5.3, why is it sensible to use TF transforms? Why not calculating a mean for the striking time range in Figure 5.3 to separate the two data sets? The answer to these questions is that the separability for a simple time domain average is less significant than the separability for the identified TF transform coefficients. This statement will be confirmed in the following.

Suppose we calculate a time average from 0.3 s until 0.7 s for each measurement of the two data sets. This yields two distributions for the calculated means referred to as $\overline{x}_{\text{Ne } 0.3/0.7}$ and $\overline{x}_{\text{Pa } 0.3/0.7}$ with a sample size of 24 each. The time frame over which is averaged is illustrated by Figure 5.8.

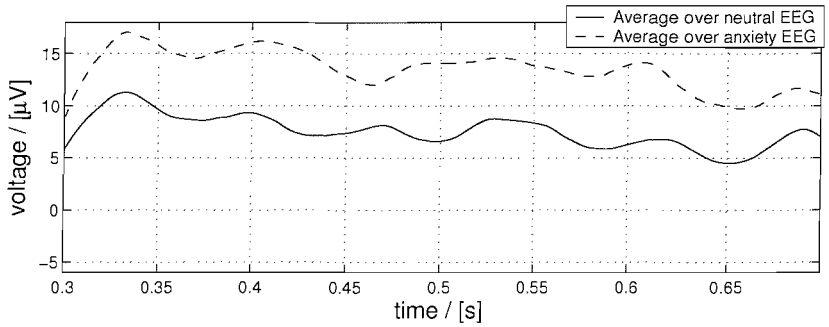

Figure 5.8: Time frame for simple time domain averages $\overline{x}_{\text{Ne } 0.3/0.7}$ and $\overline{x}_{\text{Pa } 0.3/0.7}$.

Applying a t-test to $\overline{x}_{\text{Ne } 0.3/0.7}$ and $\overline{x}_{\text{Pa } 0.3/0.7}$ yields a significance level of $P = 0.0234$. Therefore, the significance level is higher than for the selected TF coefficients to describe the difference. Moreover, if the range from 0.3 s to 0.7 s is divided into 0.02 s intervals yielding 20 intervals in total, the value of the interval with the lowest significance level equals $P = 0.0142$ which is still above the significance level of the TF transform coefficients. This shows that although relatively good values for the significance level using simple time domain averages can be achieved, the separability results for the TF transform coefficients are better.

5.3.2 Applying no or other Transforms like the DFT and KLT

To further justify the chosen transforms, we show in the following the reconstruction results for the resulting coefficients yielded when the statistical tests are applied to mere time domain data, Fourier–transformed data or data transformed with a KLT. Similar to Figure 5.7, we use the obtained coefficients to reconstruct the difference and compare them with the difference of the averages of the two data sets. Figure 5.9 shows the reconstruction results for the time domain (top) and frequency domain (middle) averaged neutral and anxiety data. The bottom of the figure represents the range between 0.25 s and 0.7 s on an enlarged scale.

We see at the top of the figure that two coefficient groups with a total of 18 coefficients at around 0.41 s and 0.6 s are identified in the time domain based on a t-test. In the frequency domain, two

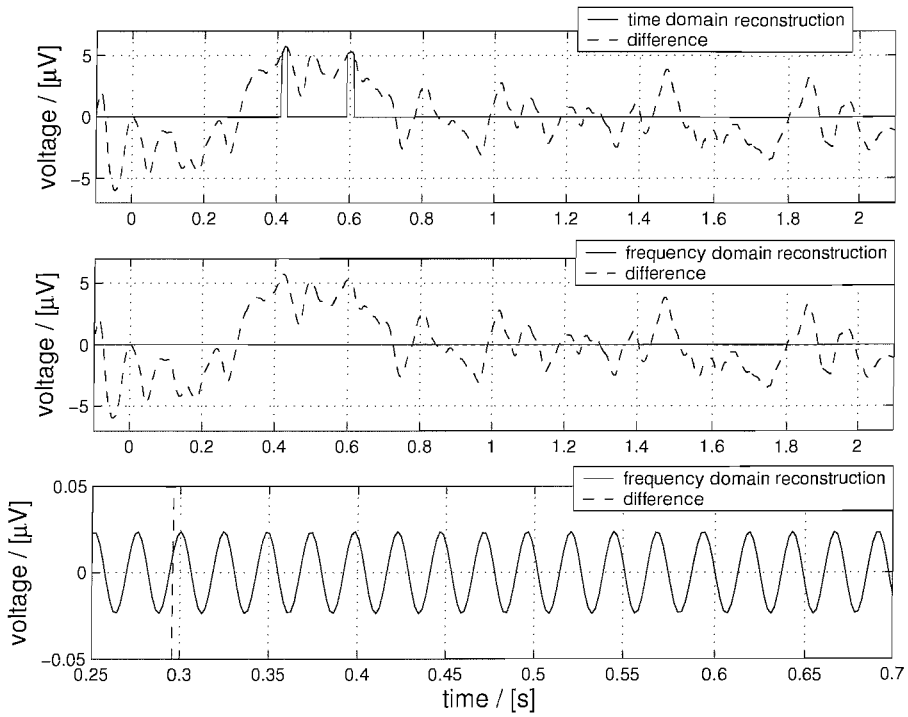

Figure 5.9: Top: difference of neutral and anxiety EEG data compared with the identified coefficients in the time domain. Centre: difference of neutral and anxiety EEG data compared with the back-transformed identified coefficients in the frequency domain (broad line at 0 μV). Bottom: enlarged scale to illustrate the approximate sine resulting from the DFT analysis.

coefficients are identified which back transformation to the time domain yields an approximate sine with a relatively small amplitude. This approximate sine cannot be clearly observed in the middle of figure, and is therefore illustrated at the bottom on an enlarged scale.

Next, the results for a parameterisation of the data by a KLT are stated. The transform is conducted as explained in 2.1.4 according to equation (2.7). In more detail, a SVD is conducted for each data set meaning that a SVD is applied twice over 24 responses for each word category. This yields a transform matrix $\mathbf{U}^T$ for each word category which is then multiplied with the respective data matrix yielding the distributions of the KLT coefficients that represent the similarity with each KLT basis function. Naturally, the best approximation of the signal information is contained by the first coefficient. The P values obtained by the t-test for the first KLT coefficients are: 0.026, 0.435 and 0.367 for the 1st, 2nd and 3rd coefficient, respectively. The reconstruction of the difference based on these coefficients is illustrated in Figure 5.10.

As it can be observed in the figure, the reconstruction of the difference for the first KLT coefficient resembles the difference very well for $t > 1$ s, but less for the interesting range from 0.3 s -0.7 s. The other presented coefficients do not show a good parameterisation of the difference either. This can be the explanation for not passing our t-test threshold of $P = 0.01$.

The above results show clearly that a parameterisation of the difference, based on the statistical tests introduced, can be much better accomplished by the introduced TF methods than with a mere parameterisation in the time domain, frequency domain or a complete adaptation to the data by a KLT transform.

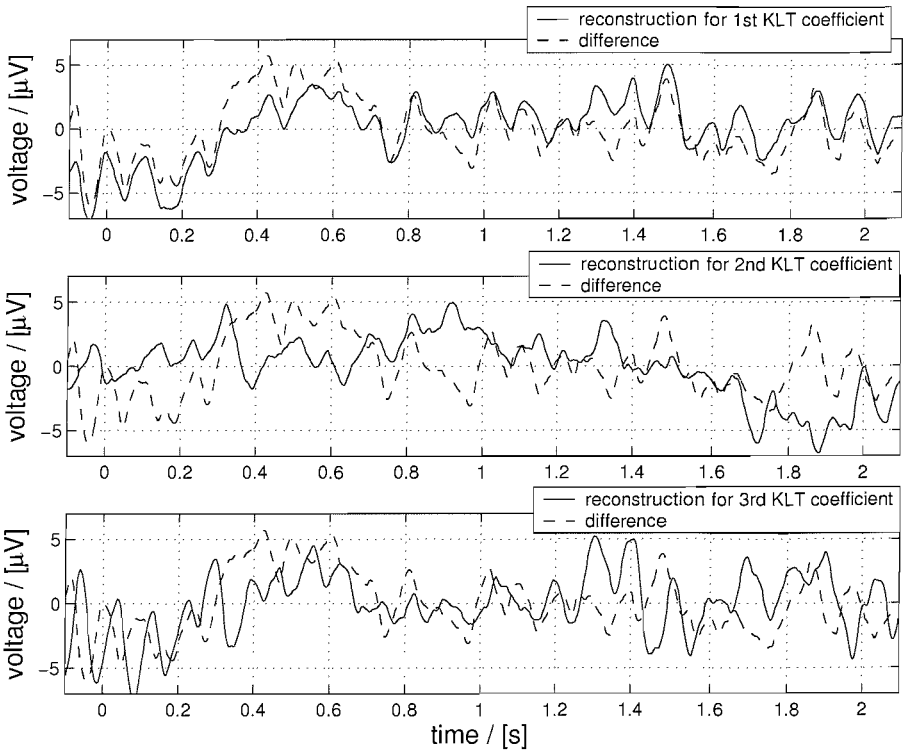

Figure 5.10: Reconstruction results for the first three KLT coefficients

5.3.3 Parameterisation Comparison for the Feature Selection Using SVM

Finally, we compare the effect of the TF parameterisation with the unparameterised case. The features that are selected by statistical tests represent the input to a SVM classifier yielding a detection rate for the TF parametrised data. This is compared with detection rates obtained for unparameterised time domain data.

When the selected features are used as an input for a SVM learning machine, it returns a trained support vector classification network. Using this network, a detection rate for a test data group is obtained. This method, as outlined in Figure 5.11 is applied to panic disorder data. We are especially interested in investigating the influence of the parameterisation and therefore, we compare the detection results obtained for the selected features based on the TF parameterisations and unparameterised time domain data. As the amount of panic disorder data is limited, the TF transforms as well as the statistical tests are applied to all available data, before it is split into training and test data sets, as illustrated in Figure 5.11 to ensure a robust parameterisation. We consider a two class classification problem, namely one class defined by anxiety causing data, and one class describing neutral data, which is split into two data groups, namely training data and test data, respectively. As the main purpose of this approach is to evaluate and justify the parameterisation, the drawback compared to a study where the parameterisation is based on a training data set only can be accepted.

In the following we describe the SVM classification in more detail.

For Gaussian kernels, when using stretched out values for the limitation of training errors defined by C and the kernel parameter γ_R, overfitting can occur meaning that all training vectors are identified as support vectors. To avoid this for our application, we have chosen to use the polynomial kernel of order $d_p = 3$ as this is assumed to be the best comprise between computational time, avoiding

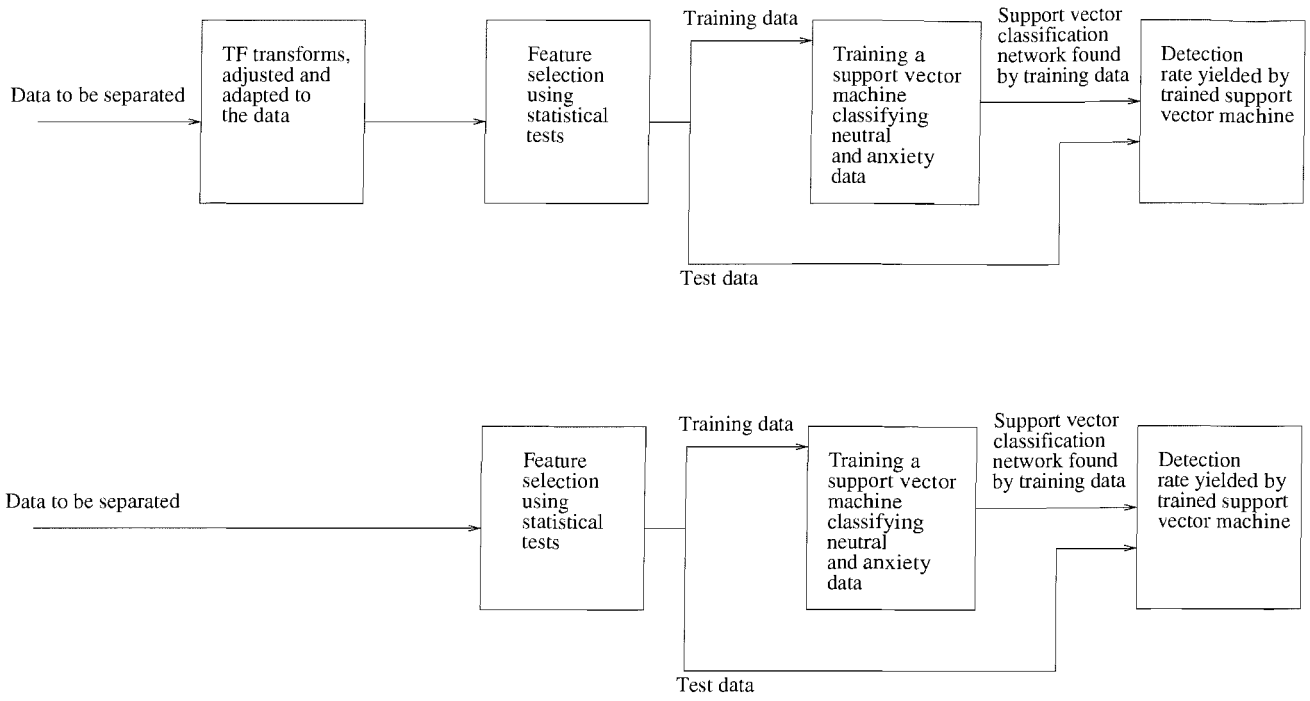

Figure 5.11: Overview: Detection comparison study for (top) parameterised data and (bottom) time domain data.

overfitting and yielding a good detection rate for the test data.

Figure 5.12 shows a SVM classification with a minimised l-o-o error for a WP parameterisation. This can be shown in a two dimensional plane as the for the WP, two coefficients were identified as significant. The two coefficients are illustrated in Figure 5.6 (left).

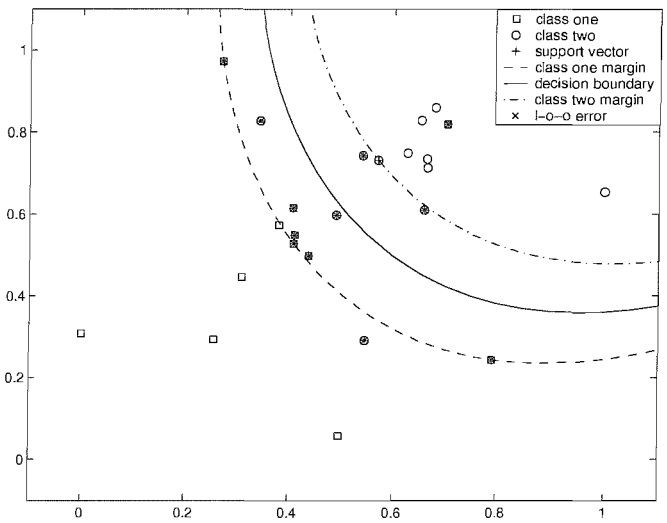

Figure 5.12: SVM classification with two coefficients for a WP parameterisation on the axis.

In Figure 5.12, one class is defined by 12 points originating from arbitrary chosen neutral words and the second class represents 12 panic causing words, also chosen arbitrary from the whole 24 defining one training data set. For this example, for class one, 11 out of 12 examples are assigned correctly;

class two gets assigned incorrectly in $\frac{1}{4}$ of all cases. This rather asymmetric decision is due to the comparably small data size. Therefore, the SVM classification was conducted with 120 samples for each word category without replacement and averaged at the end. This means that in loop of 100 runs, 12 arbitrary measurements for the two data groups were chosen, the SVM trained and the test group consisted of the remaining 12 measurements for each word category. In the next run, again 12 arbitrary measurements were chosen from the set of 24 for each category and so on and the average calculated at the end. Table 5.3 shows the results for this procedure for the training data. It is well worth mentioning that all our results are based on the implementation of the SVM according to [59].

	time domain	GF transform	WP transform
sensitivity	99.55 %	66.78 %	92.00%
specificity	93.13 %	65.88 %	96.28%
number of support vectors	15.68	16.65	17.39

Table 5.3: Results for training data.

It can be seen that the time domain data and the WP parameterisation yield comparably high values, whereas the GF identifies around two out of three words correctly for the training data. The number of support vectors is similar for each case.

Next, the more interesting results for the test groups are presented in Table 5.4. It can be seen

	time domain	GF transform	WP transform
sensitivity	67.34 %	78.49 %	82.32%
specificity	55.63%	54.39 %	52.82%

Table 5.4: Results for test data.

that the WP shows the overall best detection results for the test data, followed by the GF and the time domain data performing worst. However, for the specificity all three cases are similar around 50%. Also, for the GF, there is an improvement for the sensitivity, which can be due to the oversampled characteristic of transform.

What can be expected from Figures 5.7 and 5.9 is confirmed: The WP parameterisation yields the best detection rates, followed by the GF transform and the unparameterised time domain data performing worst. However, the specificity for the TF-transforms is not significant although the test data is used for the adjustment of the parameterisation methods. This can be due to the relatively small amount of data available. Moreover, the t-test for obtaining distinctive coefficients may not be powerful. Therefore, more encouraging results for the analysis of biomedical data applying SVM can be expected when more measurements are available than here and a different method for the extraction of the features from the parameterised data is deployed. However, it is confirmed that the parameterised data can be separated better than without a parameterisation.

5.4 Summary

We have presented a WP and Gabor transforms analysis comparison for parameterising ERP with the aim of differentiating between presented neutral and anxiety words to a subject with panic disorder. We have motivated the use of TF methods, and proposed an approach to obtain distinctive transform coefficients, whereby the results were verified by different tests for different cases. The selection of those transforms was confirmed by a comparison with other parameterising transforms, namely the DFT and the KLT. Moreover, a comparison with a simple time domain average and a test with completely unparameterised data was conducted. The results of those comparisons are that only considering the time domain data or applying a DFT is too "inflexible" to achieve good results as no adjustment to the data can be done. However, a complete adjustment to the characteristics of the data by a KLT does not yield satisfying results either. Also, calculating a simple time domain average does not outperform the separability results for the TF parameterised data. As the author is not aware of other work analysing panic disorder by TF transforms, these justifications shall confirm the described approaches. Therefore, it is concluded that these TF transforms can be very well used to analyse and select the features of panic disorder via ERP responses for the study shown.

Chapter 6

TF Transform Based Detection of Cochlear Hearing Loss Using SVM

This chapter is concerned with the automated separation of biomedical data to detect and specify potential cochlear hearing loss by means of transient evoked otoacoustic emissions (TEOAE) by combining the parameterisation methods of Chapters 2 and 3 with the classification in Chapter 4. In more detail, we give an overview of otoacoustic emissions and their application to medical diagnosis in Section 6.1. Then, we describe and discuss the results for each of the studied parameterisation methods showing the DWT in Section 6.2, the WP in Section 6.3 and the GF transform based analysis results in Section 6.4. Each of these sections has the same layout, comprising of feature extraction, feature selection and classification. We compare and discuss the results including a comparison to a similar study in Section 6.5. Finally, the chapter concludes with a summary in Section 6.6.

6.1 Introduction

This section gives an overview over otoacoustic emissions and a brief review on screening and diagnosis methods based on these signals.

6.1.1 Otoacoustic Emissions

The current understanding of sound processing in the inner ear includes two components: a passive and an active mechanism. Interaction of these two mechanisms gives the inner ear its extremely high sensitivity and frequency-selectivity. The active amplification of the displacement of the basilar membrane produces vibrations which are transmitted backwards through the ossicular chain in the outer ear canal. Thus the inner ear does not only transform an acoustic stimulus into an electrical signal but can also actively emit sound. These sound emissions are considered as an epiphenomenon of the active cochlear amplification process and are called otoacoustic emissions (OAE) [2]. OAE are transmitted backwards to the physiological direction of sound into the outer ear canal where they can be measured by highly sensitive microphones.

Basically we can distinguish between spontaneous (SOAE) and evoked otoacoustic emissions

(EOAE). EOAE were first systematically described and analysed by Kemp [68], and are classified according to the applied stimulus into

- transiently evoked emissions (TEOAE),
- stimulus-frequency emissions (SFOAE) and
- distortion product emissions (DPOAE).

It is beyond controversy that the spontaneous and the evoked emissions originate in the cochlear amplifier and that they both occur only in a healthy inner ear [2],[7],[6]. In ears with a hearing loss of more than 35 dB it is impossible to detect any TEOAE [2]. However this is a frequency-specific effect; if there is a hearing loss in a strictly limited frequency range we can still measure emissions in the same ear in adjacent frequency ranges with normal hearing.

The TEOAE are described in more detail next.

6.1.2 Transient Evoked Otoacoustic Emissions (TEOAE) and Their Measurement

TEOAE are evoked by brief acoustic stimuli, for instance a click (1-6 kHz) or a tone-burst. For the measurement an acoustic probe containing a loudspeaker for the delivery of the stimulus and a sensitive miniature microphone for the recording of the emitted sound signal are inserted into the external ear canal. Broadband acoustic signals from the cochlea are recorded when the inner ear is stimulated with a click. With a tone-burst the registered emissions correspond to the frequency of the stimulus. This allows to a certain extent a frequency-specific statement on the intactness of the outer hair cells inside the cochlea [2].

After a frequency-dependent latency the evoked sound emissions can be measured in the outer ear canal. Because of the frequency dispersion due to the representation of high frequencies at the base of the cochlea and of low frequencies at the apex, the latencies are strongly dependent on the emission frequency, e.g. 20 ms at 500 Hz and 4 ms at 5 kHz are common [2]; alternatively we can say that cochlear answers occur with a frequency-dependent time delay. The duration of the TEOAE ranges from a few ms to several hundred ms.

An important characteristic of TEOAE is their non-linear behaviour: the amplitude increases up to stimulus levels of 35 dB sound pressure level (SPL) and saturates at higher stimulus levels. This non-linearity is important in the signal processing and evaluation stages. State-to-the-art measurement equipment [52] offers a non-linear stimulus mode, which releases a stimulus sequence of three positive acoustic impulses, referred to as "click" stimulus, followed by a three times greater one possessing an opposite phase meaning that it has a negative sign. The sum of the responses to this stimulus set is stored. This stimulus mode results in an average response that reduces stimulus artefacts and linear components in the response of the ear to the transients [2].

Standard equipment such as in [52] measures TEOAE synchronously to the stimulus over a certain time interval, and its time series is represented in a vector $\mathbf{x}$. As TEOAE data is prone to a considerable contamination by noise, the measurement equipment performs averaging over a total of L responses to the above described stimulus set, whereby two partial averages, $\overline{\mathbf{x}}_{\mathrm{A}}$ and $\overline{\mathbf{x}}_{\mathrm{B}}$, are formed according

to (3.11) and (3.12). Per measured ear, stimulus synchronous intervals of 20.48 ms represented in samples indexed by n are averaged with a sampling frequency of $f_s = 25$ kHz. The number of averaged stimulus set responses is 260 for each average corresponding to $L = 520$ in default mode for [52]; the sample number equals $N = 512$. As there are 4 stimuli for each stimulus set, the two averages contain the responses to 2080 stimuli for one ear.

The measurement of TEOAE with the ILO88 measurement equipment [52] is part of the clinical routine to check the hearing ability of newborns. To describe the measurement procedure in more detail, Figure 6.1 shows a TEOAE measurement of the author. The applied "click" stimuli excite the frequency range from 0 Hz up to 6 kHz in the cochlear.

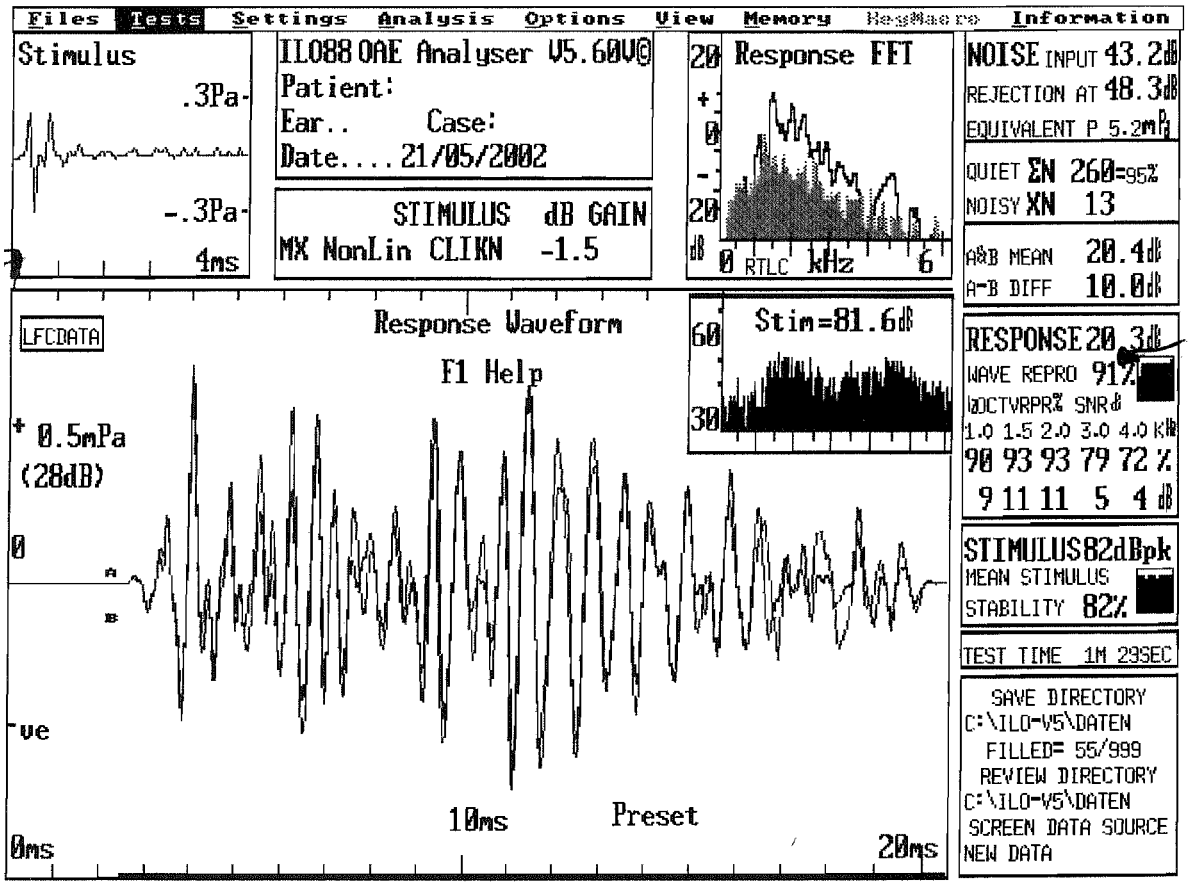

Figure 6.1: TEOAE measurement of the author with ILO88 measurement equipment.

The response waveform shows one partial average over 260 stimulus sets. Among the illustrated measurement results in Figure 6.1, the SNR according to equation (3.14) can be determined as the difference between the A&B mean and A–B diff equalling 20.4 dB−10.0 dB = 10.4 dB for the author.

The most important measurement for the operator is the SNR value for the different centre frequencies of 1.0, 1.5, 2.0, 3.0 and 4.0 kHz. The values are illustrated in the window titled response, below the correlation coefficient values which are given in %. Values below a threshold of $\approx 5-6$ dB are critical indicating a possible hearing loss. In this case, a clinical operator might diagnose a slight high frequency hearing loss for the author.

To illustrate the differences for subjects with different hearing ability, Figure 6.2 gives an example

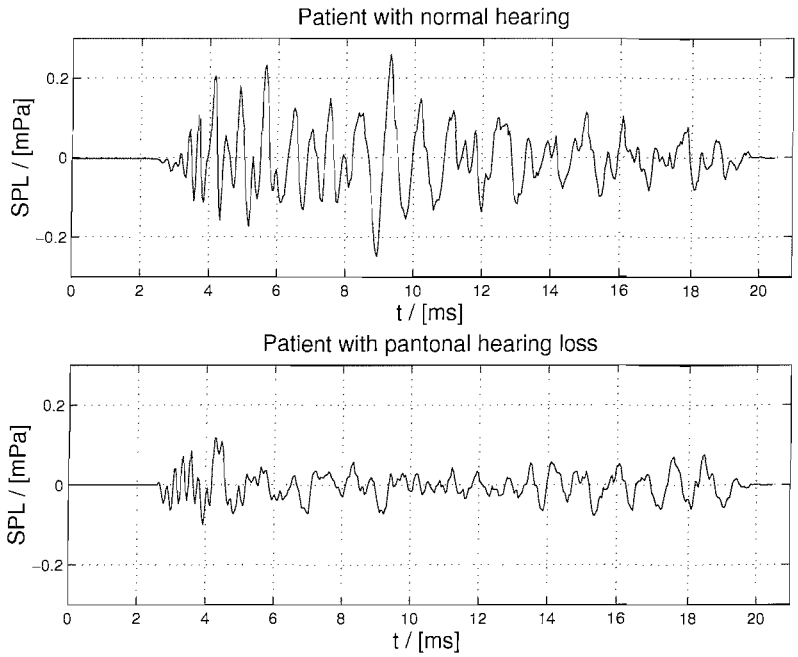

Figure 6.2: TEOAE for subjects with (top) normal hearing ability and (bottom) pantonal hearing loss.

for a TEOAE signal measured by [52] which shows the recorded partial average $\overline{x}_A$ for a subject with normal hearing ability on the top and the recorded partial average $\overline{x}_A$ for a patient with pantonal hearing loss on the bottom.

TEOAE can be measured in more than 90% of all subjects with normal hearing. This important feature contributes to the clinical relevance of TEOAE. In subjects with normal or almost normal hearing (hearing loss up to 15-20 dB) the cochlear answer registered after click stimulation is a broad-band frequency spectrum. In contrast a hearing loss of more than 35 dB leads to a complete loss of active sound emissions. A localised damage of cochlear hair cells results in a reproducible reduction or even total loss of TEOAE in the corresponding frequency ranges [2].

The measurement of TEOAE is established in the clinical routine as an objective and non-invasive diagnostic tool. The measurement of TEOAE is very fast and brings minimal strain for the subject and is therefore an excellent screening method for the detection of cochlear hearing disorders in newborns and infants. Especially the early diagnosis of hearing impairment in children is most important for a prompt initiation of an adequate therapy which can avoid irreversible damage.

After introducing TEOAE in general, the data used for our studies is described.

6.1.3 Description of Data and Standard Analysis

We try to differentiate between three different types of hearing loss (HL) – no HL, high-frequency HL, and pantonal HL as characterised in Figure 6.3 where the shaded area specifies a possible HL.

The available TEOAE data consists of two data sets recorded at the Universities of Homburg and Heidelberg in Germany with measurement equipment according to [52] as described in Section 6.1.2, whereby each set evaluated roughly 200 ears, the exact numbers are given in Table 6.1.

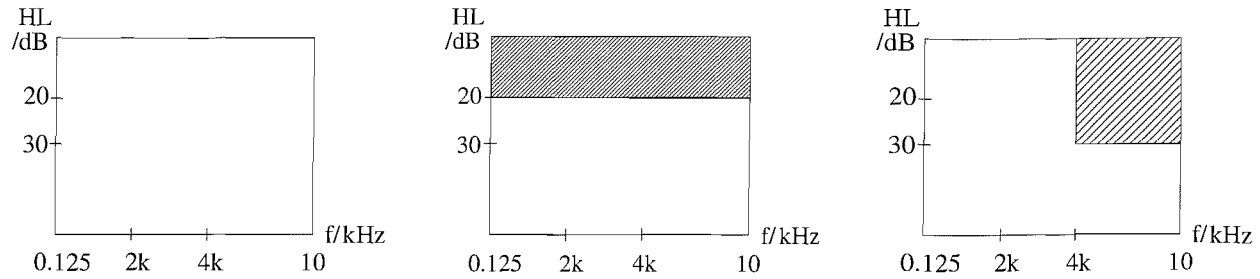

Figure 6.3: Characterisation of hearing loss for (left) normal hearing, (middle) pantonal HL, and (right) high frequency HL. The shaded area indicates a possible hearing impairment.

number of subjects	Homburg	Heidelberg
NH	109	69
HF	55	95
PT	23	39

Table 6.1: Number of subjects for each hearing ability group for each data set.

The measurements represent the response to a "click" stimulus as shown in Figure 6.1 on the top left. As stated before, each stimulus set consists of 3 stimuli as shown in the figure, and a fourth one with a three times bigger amplitude and an opposite sign to reduce linear components. For each partial average 260 stimulus set responses are averaged equalling 2080 stimuli for one ear. The medical history of the subjects in the data sets was known, and clear assignments to one of the three groups of different hearing ability, as defined in Figure 6.3, had been established at the Homburg and Heidelberg clinics. The author would like to acknowledge Prof. Ulrich Hoppe and Sebastian Hoth of the University of Erlangen, Germany, who kindly provided valuable expertise and the data.

In the following we will show the results for an ROC analysis of the standard SNR values of the data sets. The SNR calculation according to equation (3.14) yields distributions for the SNR for each hearing ability group for the two data sets. E.g. Figure 6.4 shows the histogram for the SNR for the NH group of the Homburg data set.

Based on the SNR distributions, we can calculate the area under the ROC curve for each distinction case for the two data sets. Table 6.2 shows the results.

group distinction	Homburg	Heidelberg
NH — HF	0.772	0.773
NH — PT	0.856	0.953
HF — PT	0.668	0.840

Table 6.2: ROC area values based on SNR distributions.

It can be seen that the Heidelberg data is of better quality than the Homburg data. The distinction case NH vs PT is the easiest to separate, while the case HF vs PT is the most difficult. For the case NH vs HF, both data sets show the same ROC value. With our signal processing methods to be described in later sections of this chapter, we aim at improving these separabilities.

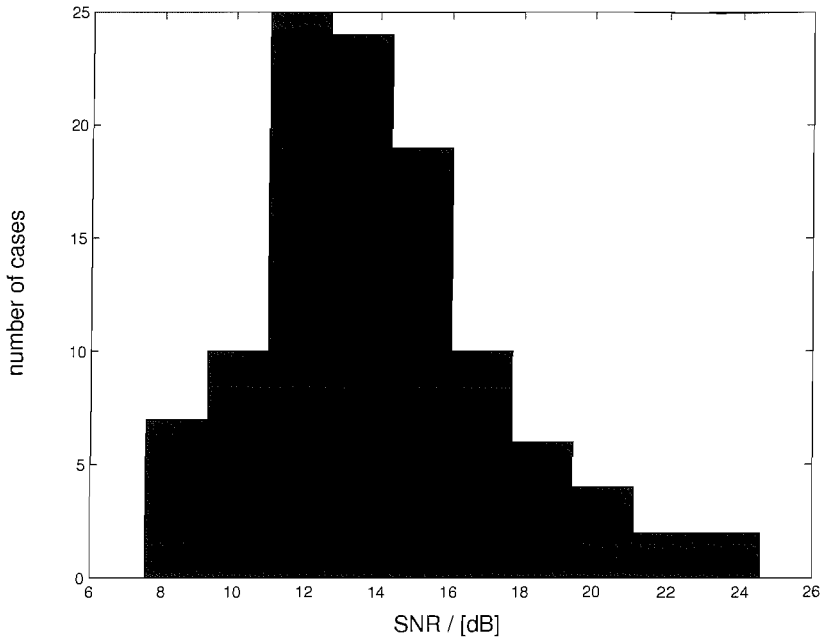

Figure 6.4: SNR histogram of the NH group of the Homburg data.

6.1.4 TEOAE Analysis Comprising TF Transforms

The tonotopic arrangement of the hair cells and the temporal aspects of the generation and propagation of anterograde and retrograde travelling waves account for a differently delayed appearance of TEOAE parts that originate at different locations on the basilar membrane [69]. Consequently, the spectrum of TEOAE signals is latency-dependent [70, 71], whereby early emissions have a higher frequency than the ones that appear with an extended latency. Thus, time-frequency methods have been considered in order to better exploit this TEOAE characteristic and gain access to the different spectral component of the TEOAE bearing evidence of frequency dependent hearing.

There exist quite a few studies on TEOAE. In [72], a "banana-like" TF pattern is identified for TEOAE via STFT and Gabor analysis based on Gaussian shape windows. This pattern is also obtained when a DWT based on Mallat's mother wavelet is applied to measured partial averages for a group of normal hearing subjects as shown in Figure 6.5.

The study [72] also compares the TF analysis for people with normal hearing ability and patients with moderate and severe high-frequency hearing loss. In the latter, high-frequencies are absent in the TF plane and the frequency interactions at later TEOAE time segments decrease. Also, for a differentiation of normal hearing with severe high-frequency hearing loss, all frequencies are significant. For a separation of normal hearing and moderate high-frequency hearing loss, only high frequencies are useful. For the separation case moderate high-frequency hearing loss versus severe high-frequency hearing loss, low frequencies are significant.

Further studies deal with enhancing the TEOAE measurement using filtering. E.g. in [73], the detection of TEOAE is enhanced by low-pass filtering leading to a better measurement efficiency. It is shown, that the same SNR value can be achieved by averaging 60 filtered sweeps as compared to 260 averaged sweeps without filtering. According to [74], using a shorter time window from 2.5 ms to 7.5 ms or 9 ms enhances TEOAE measurement efficiency, which results in a greater SNR. The usual

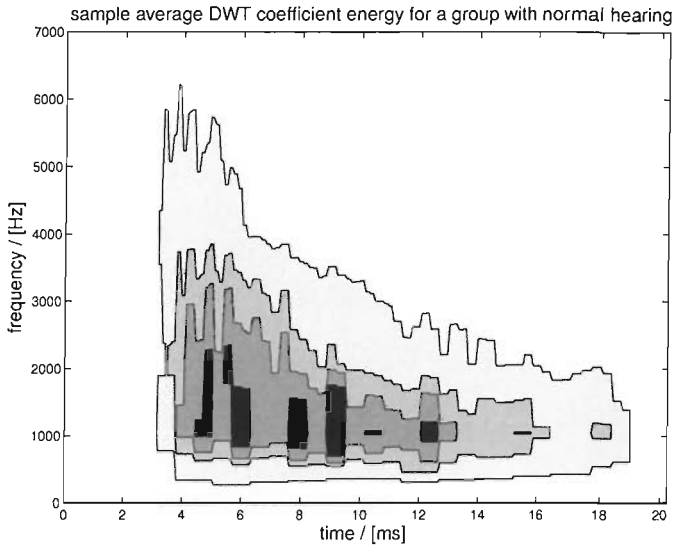

Figure 6.5: TEOAE energy in the TF plane showing a "banana" pattern.

window length is up to 20.48 ms as mentioned above. More responses can be averaged in the same time by using the shorter 2.5 ms to 7.5 ms time windows, such that a specified SNR can be reached up to five times faster.

Wavelet-type analysis of TEOAE is presented in [75] to study the influence of a drug, using a 5 tap finite impulse response (FIR) filter as a prototype in a filter bank. The study concludes that TEOAE can be suppressed by drugs. Also, [76] includes the analysis of TEOAE by wavelet transforms, where the mother wavelet is a modulated cosine function. However it can be questioned if such a function can be attributed as a mother wavelet for a DWT. The study uses this type as mother wavelet and confirms the above mentioned "banana" pattern by revealing linearity in a logarithmic scale for latency versus frequency. Furthermore, a monotonic decrease is found in the latency-frequency diagram in logarithmic scale for increasing the stimulus intensity meaning that all frequencies appear faster for a higher stimulus. Pseudo Wigner Distributions are used to uncover the TEOAE characteristics in the TF domain in [77], indicating another confirmation of the "banana" pattern.

Furthermore, matching pursuit can be applied for TEOAE analysis [78]. It is claimed to provide a higher resolution than the DWT or WP. By overlapping 20 functions adequately chosen from a redundant dictionary, 80% of the energy can be represented. Also, a relation between the stimulus level and the frequency components is uncovered: For increasing stimulus, the level of the high-frequency components increases significantly less than the low frequency levels.

In [6] different tests for the objective detection of TEOAE are described. Standard measurement equipment [52] gives parameters like an SNR value and a correlation coefficient. Among others, these two parameters are also included in the study. The conclusion is drawn, that a variance ratio test is the best suited for screening and detection applications, better than the SNR estimate or the correlation coefficient. This shows, that there is space for improvement over obtained standard results from the measurement equipment.

A differential diagnosis of hearing loss based on TEOAE is conducted in [7]. There, a design of time windows is shown, which increases the separation results compared to the unwindowed case by

15%. In more detail, a parameterisation of the data by wavelet decomposition using the Coiflet5 as mother wavelet is conducted prior to applying an ensemble correlation technique. This is followed by deploying the time window design. Finally, using a mean cross-correlation, the improved separability values are achieved. Concerning our study, those results can be used for comparison.

Having introduced TEOAE and discussed their clinical application, a novel method is developed for the differential diagnosis of hearing loss based on TEOAE in the next sections. We start by giving an outline of our approach next.

6.1.5 Overview of our Analysis Approach

We aim to improve the results shown in Table 6.2 by employing a range of TF parameterisations, apply the feature selection shown in 3.4 and classify the data with SVM according to 4.3. As the TEOAE data is not overlapping, the procedure described in Section 3.4.2 is left for future applications.

In recent studies [79], [80] a SVM classification was conducted after signal processing methods were applied for the artifact removal of EEG data. Hence, the application of a SVM after the described feature extraction and selection seems also reasonable for TEOAE. Moreover, the application of the SVM classifier after our signal processing methods is justified by the results shown in Subsection 5.3.3. There, the time domain based SVM classification yielded the least significant results even though a feature selection method was applied. Therefore, it can be expected that using the pure time domain TEOAE data as input for a SVM will lead to similar results even though the dimension of the classifier is larger than the number of points defining the classes. For example in Section 5.3.3, 18 time domain coefficients were used to classify 24 data points. Applying no feature selection to the data with a sampling number of 266 yielded even poorer results there.

Based on these explanations, we arrive at the following overview of our analysis approach: Figure 6.6 shows that a ROC analysis after the feature selection and the classification is conducted in order to compare the results with Table 6.2. For the SVM classification, the results for a DAGSVM test and a DAGSVM test with a neutral class will be given. The latter classification will be explained and determined next.

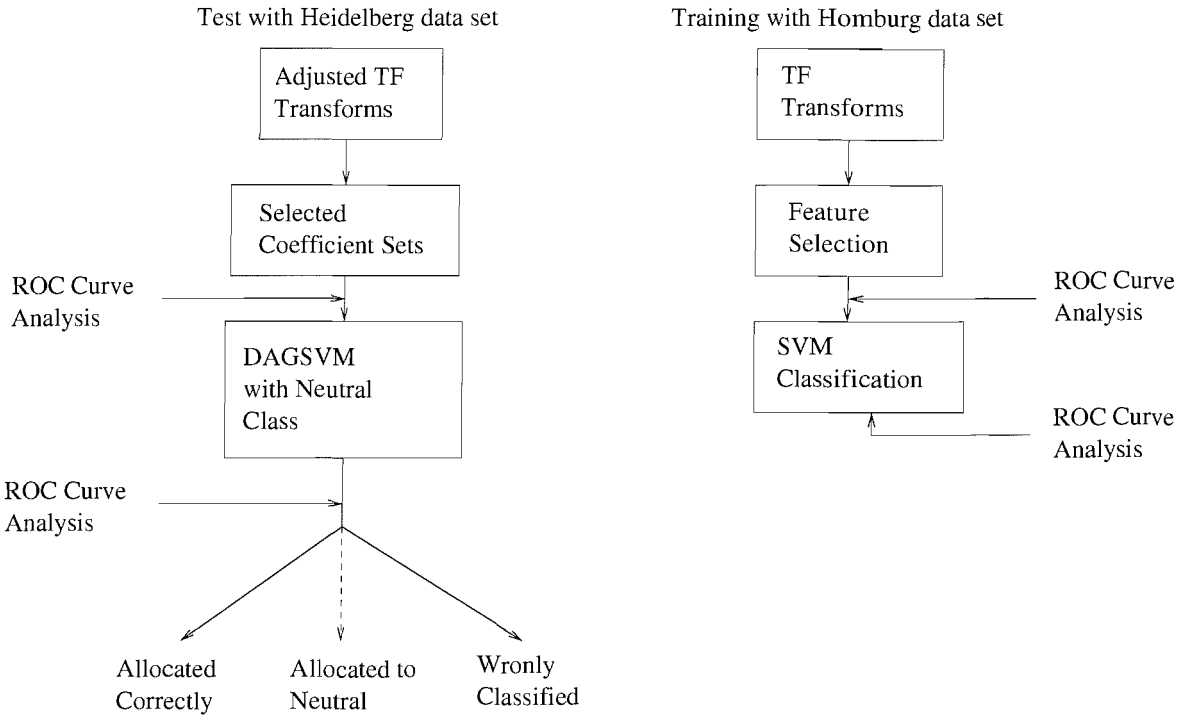

Figure 6.6: Overview of detection approach for cochlear hearing loss.

As Table 6.2 shows, the case NH vs PT is the easiest to separate and hence, we determine that a SVM classification without a neutral class can be applied for this case. For the cases NH vs HF and HF vs PT, which are more difficult to separate, we define a neutral class as explained in Section 4.3.4 with maximal possible margins to illustrate the extreme case. The DAGSVM decision tree for this case is shown in Figure 6.7 which is based on a separation of the classes according to Figure 6.8.

The reason for this determination of the DAGSVM with a neutral class is that if we applied a DAGSVM decision as illustrated in Figure 4.21, the HF class would be required to be classified as neutral first and then pass another two tests. As the NH vs PT classification is based on a specially adapted coefficient set for this case, the outcome when testing the HF class seems quite unforeseeable. Therefore, it makes sense to use a hard decision here, where the HF class is divided up at the first classification node. For the following decisions the neutral class can contribute to assure a robust classification as the HF group is more difficult to separate from the other two.

With these explanations, we give the results for the analysis for each parameterisation method in the following sections. All results shown within this chapter are based on the implementation of SVM according to [59] applying a Gaussian kernel.

6.2 DWT Analysis of TEOAE

This section presents the results for the detection of cochlear hearing loss based on TEOAE for a DWT parameterisation of the data. The section is divided in the three parts: feature extraction, feature selection and classification. The results are evaluated with a ROC which can be compared with the standard analysis results in Table 6.2.

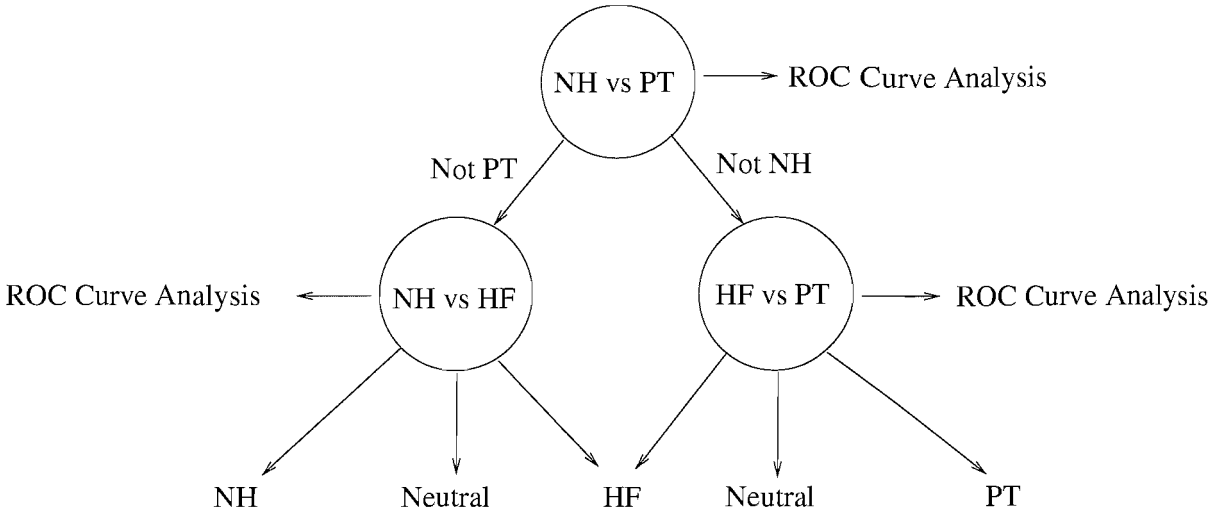

Figure 6.7: DAGSVM decision tree for TEOAE analysis.

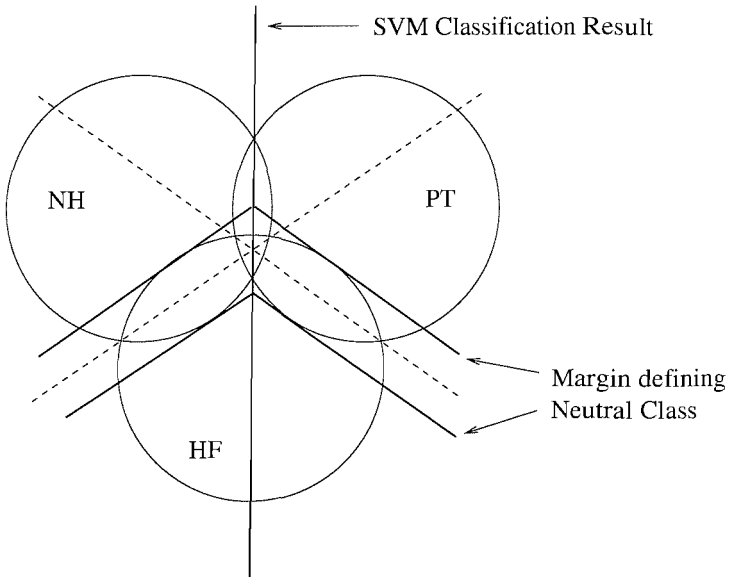

Figure 6.8: Definition of classification for TEOAE illustrating neutral class.

6.2.1 Feature Extraction: Transform Adjustment and DWT Decomposition

According to a large number of studies in the literature [81],[34],[13],[78], Mallat's wavelet [30] is well suited for the analysis of biomedical data such as EEG. Furthermore, it is symmetric, and hence allows a symmetric extension of the data. As stated in Chapter 2, periodic extension suffers from blurring features hidden close to the interval margins of the data and therefore, wavelets that can only be extended periodically are not applied. For these reasons Mallat's wavelet is used for our DWT. The impulse and magnitude response of a basis function derived form Mallat's wavelet is shown in Figure 3.3.

The energy of TF transformed data vector $\mathbf{y}$ according to (3.13) averaged over the partial DWT coefficients $\mathbf{y}_A$ and $\mathbf{y}_B$ for the three hearing ability groups for the Homburg data is shown in Figure 6.9 to indicate how the energy of a TEOAE is resolved in the TF-plane. As the figure illustrates, the PT data is fully contained in the HF data, and the HF is again fully contained in the NH data.

Therefore, we can only apply the difference evaluation method described in Section 3.4.2 because there is no coefficient set that could be identified by (3.26). The clinical interpretation of the graphs

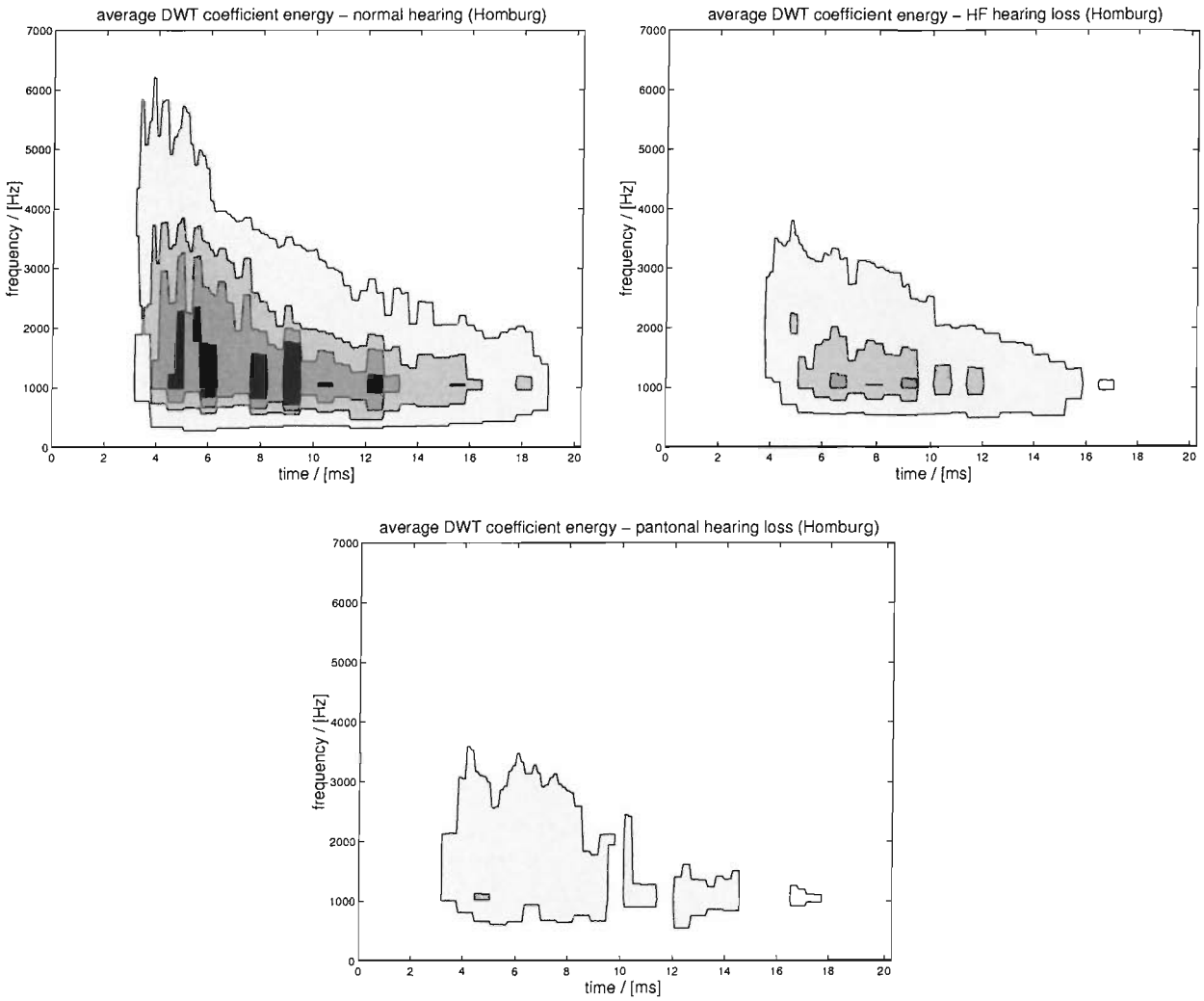

Figure 6.9: Average DWT coefficient energy for the Homburg data: (top left) normal hearing (NH), (top right) high frequency hearing loss (HF), and (bottom) pantonal hearing loss (PT).

is as follows: The spectrum of the TEOAE for NH subjects shows early high frequencies and late low frequencies that can be observed quite clearly. In contrast, the TEOAE spectrum for HF patients lacks the high frequency components found in NH subjects. Finally, the TEOAE spectrum for PT patients shows very low energy components in the whole TF plane. Hence, a good and meaningful distinction between the three hearing groups is indicated using the DWT coefficients.

6.2.2 Feature Selection

With the difference evaluation method introduced in Section 3.4.1, the resulting optimised coefficient set is characterised by their TF-coordinates in Figure 6.10. The results shown there are optimal in the sense that the best results for the Homburg data were attained.

The arrangement of the coefficients in the TF plane in the figure is reasonable in terms of their physiological meaning. Referring to Figure 6.9 it makes sense that the separation coefficients for NH

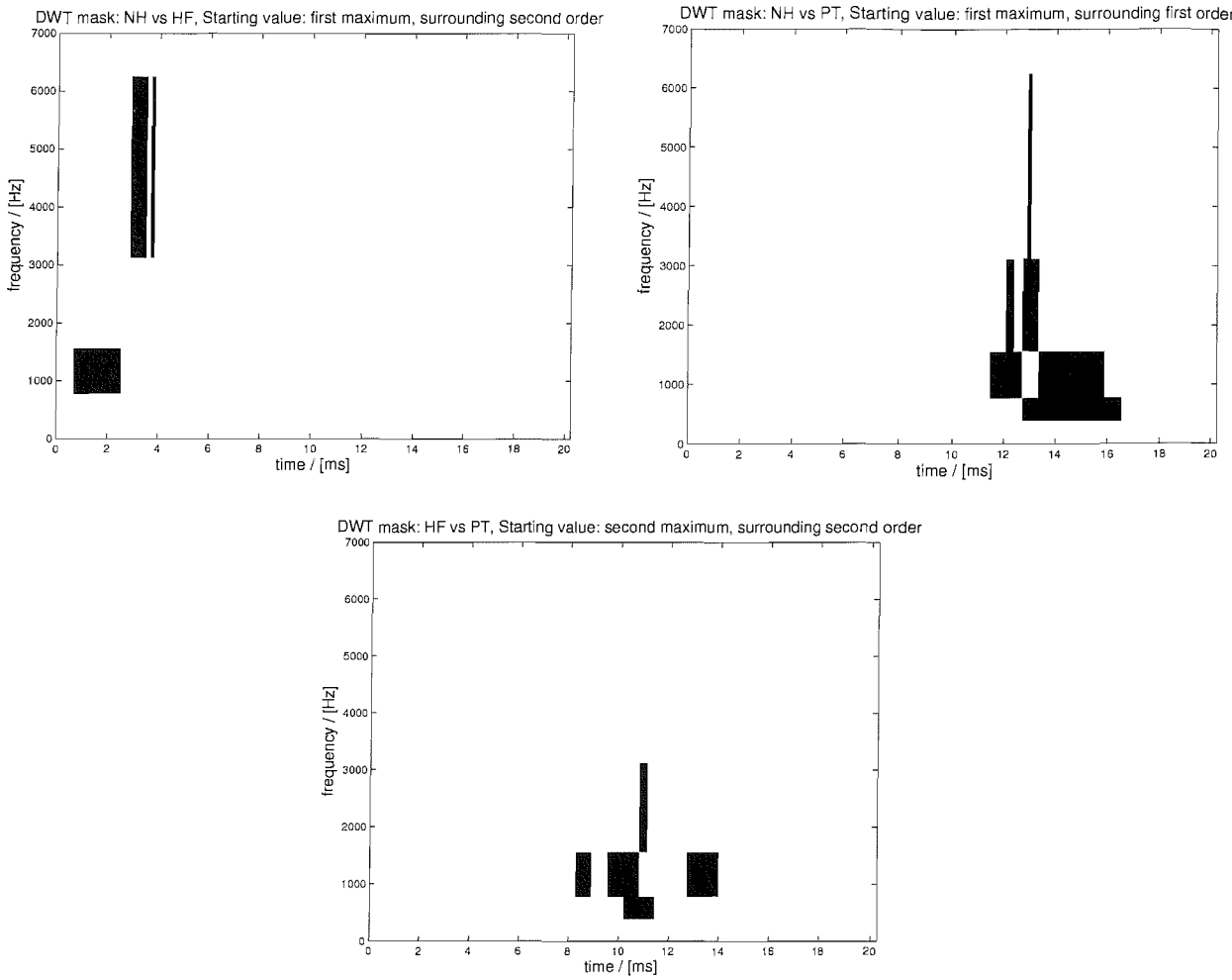

Figure 6.10: Resulting DWT coefficients in the TF plane: (top left) NH vs HF yielding 8 coefficients, (top right) NH vs PT yielding 13 coefficients, and (bottom) HF vs PT yielding 7 coefficients.

vs HF are at early high frequencies which are likely to be missing in patients with HF hearing loss. For the case NH vs PT the coefficients are located late in time and in the mid frequency range. As the average coefficient energy is quite small in that area for PT compared to NH, this result is reasonable. A very similar set of distinctive coefficients is identified for HF vs PT, which is likely to be due to the small amount of energy for PT in that area.

The resulting coefficient set is obtained by the iterative search as described in section 3.4.1. However, to get the best separability results, an exhaustive search approach would be well suited. But this, as shown in Section 3.4.3.2, is adaptable to noise and hence, it is taken out of consideration. Moreover, the results of the extensive simulations concerning the first and second order neighbourhood search show that for the case NH vs PT, the TF coefficients are too adapted to the Homburg data. As this case is the easiest to separate anyway, the search area is restricted to first order only.

According to Figure 6.10, for the distinction cases NH vs HF and HF vs PT the best separability is obtained by searching the second order neighbourhood. For the case NH vs HF, the first maximum was identified and absorbed in the resulting coefficient set. However, for HF vs PT, the first maximum is not part of the identified coefficient set. For NH vs PT, the optimised separation is achieved by starting with the first maximum and considering first order neighbourhood only for growing the coefficient set

as mentioned above.

In order to ensure that the selected coefficient set is generalisable from the Homburg data, from which it has been derived, we apply the same C_{opt} to separate between the three cases of hearing ability for the Heidelberg data. The area under the ROC curves are summarised in Table 6.3 to quantify the separability. Compared with Table 6.2 the conducted feature selection method could

	separability results by coefficient set in Figure 6.10	
group distinction	Homburg	Heidelberg
NH — HF	0.867	0.779
NH — PT	0.959	0.955
HF — PT	0.862	0.875

Table 6.3: Separability (area under ROC curve) for the 3 hearing ability groups for the Homburg and Heidelberg data for DWT.

improve the results for all cases except the HF group of the Heidelberg test data. Hence, we apply further methods like SVM aiming at further improvement.

6.2.3 Classification

For the classification, we compare two methods: A DAGSVM classification according to [62] and a DAGSVM with a neutral class as shown in Figure 6.7. Figures 6.11–6.13 illustrate the resulting ROC curves for the SVM analysis.

The figures contain graphs for ROC curves for the separability for two unit variance Gaussian distributions separated by d. As mentioned in Section 3.1, for $d = 0$, both distributions match, while for $d = 1$ and $d = 2$, their means are separated by the standard deviation or twice this value, respectively. The dots in the curves for the NH vs HF and HF vs PT node state the thresholds for the neutral class.

Observing the values of the area under the ROC curve which are shown on the top of each diagram in Figures 6.11–6.13, for the training data, a good classification is obtained for all cases which is not surprising for a SVM learning machine. Also, the NH vs PT case can be separated very well for the test data. However, there is no improvement compared to the separability value for the feature selection. For the other cases, the separability values even decrease compared to the feature selection. However, for this comparison one has to keep in mind that the HF group is split by the first node of the DAGSVM which means that for the classification, the NH vs HF and HF vs PT cases contain a divided HF group in contrast to the feature selection where the whole HF group was used.

For the normal DAGSVM, 79.7% was the detection rate for the NH test group, 63.2% for the HF group and 69.3% for the PT group which matches with Table 6.2 where it can be expected that the NH group is the easiest to detect.

The results for applying the neutral class are shown in Table 6.4. Illustratively, the neutral class can be observed in Figures 6.12 and 6.13 as the areas between the two dots on the ROC curves. We

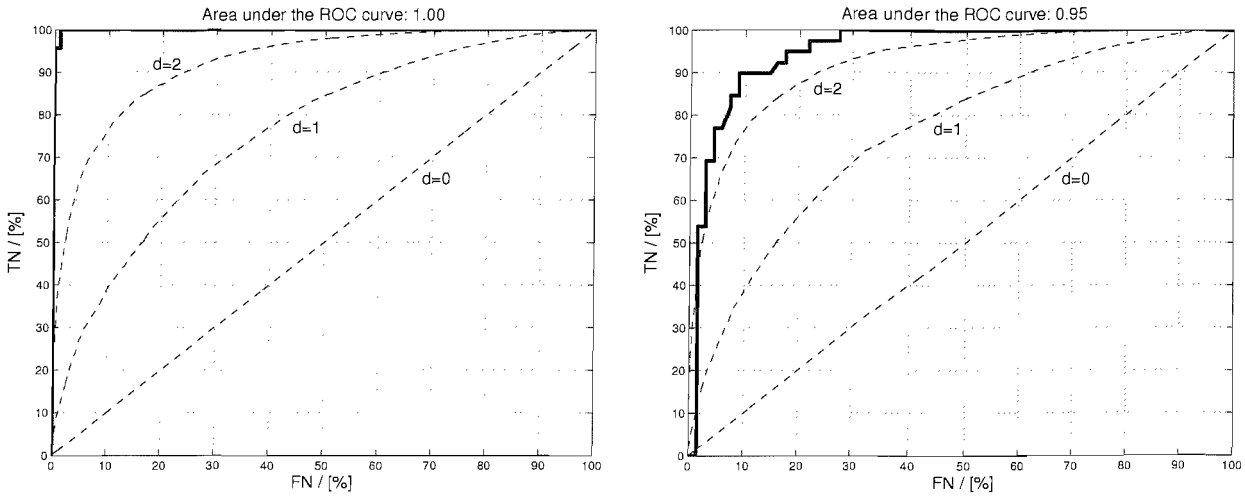

Figure 6.11: ROC curves for the NH vs PT DAGSVM node for (left) training and (right) testing based on a DWT parameterisation.

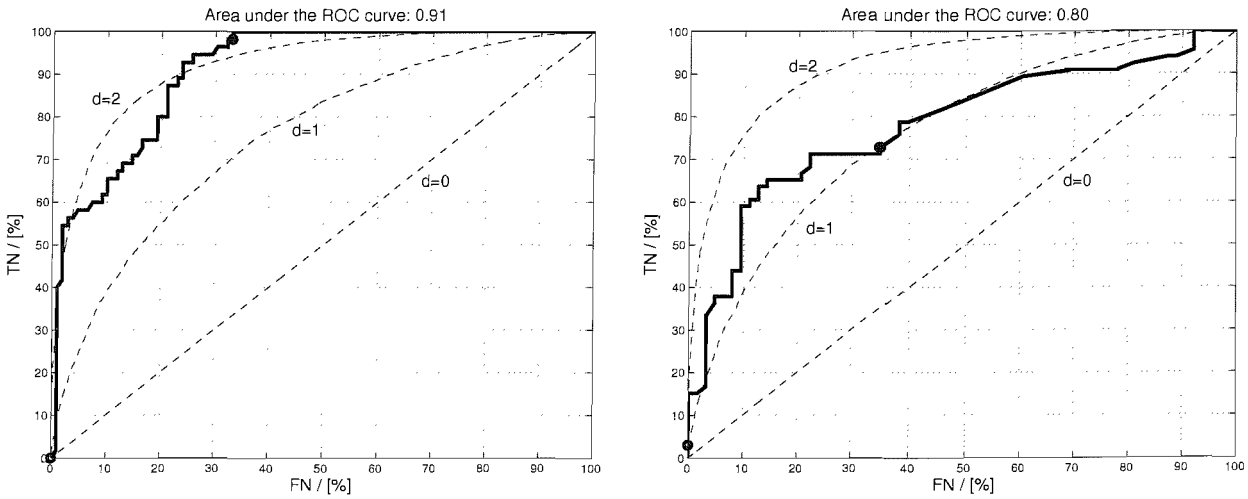

Figure 6.12: ROC curves for the NH vs HF DAGSVM node for (left) training and (right) testing based on a DWT parameterisation. The dots indicate the margins for the determined neutral class.

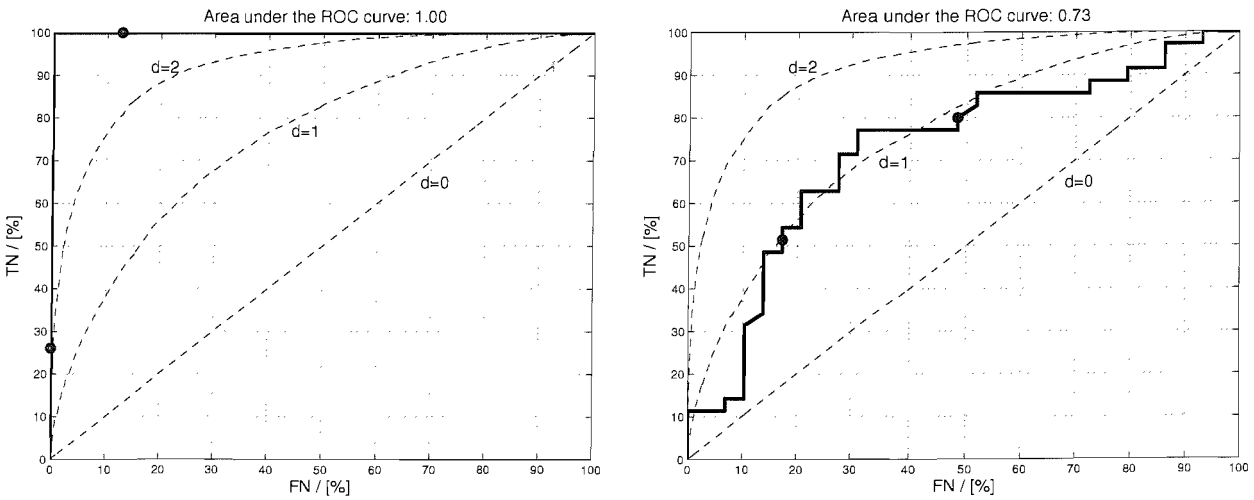

Figure 6.13: ROC curves for the HF vs PT DAGSVM node for (left) training and (right) testing based on a DWT parameterisation. The dots indicate the margins for the determined neutral class.

NH false	NH neutral	NH correct	HF false	HF neutral	HF correct	PT false	PT neutral	PT correct
8.7%	31.9%	59.4%	24.3%	57.9%	17.8%	28.3%	25.6%	46.1%

Table 6.4: Detection rates yielded by DAGSVM classification with a neutral class for test data for DWT parameterisation.

see that the NH group can be detected most significantly whereas the PT group is the most difficult to detect. For the HF group, most patients are allocated as neutral.

Comparison between the DAGSVM with and without neutral class yields that for the NH group the falsely classified rate is lowered from 20.3% to 8.7% at the cost of a decrease in the correctly classified from $79.7\% - 59.4\% = 20.3\%$ which can be viewed as acceptable. However, for the PT group, the falsely classified rate is only decreased by 2.4% with the neutral class approach along with a decrease for the correctly classified of 23.2%.

6.3 WP Analysis of TEOAE

The WP analysis of the TEOAE data shown in the following is based on the DWT implementation in Section 6.2, e.g. using the same mother wavelet. It utilises the method described in Section 2.4 to find the optimal decomposition. Again, the section is divided in the decomposition of the data, feature selection and classification.

6.3.1 Feature Extraction: Transform Adjustment and WP Decomposition

To conduct the WP analysis according to Section 2.4, we define the data matrix $\mathbf{X}_{\text{data}}$ as

$$\mathbf{X}_{\text{data}}^{L\times N} = \begin{bmatrix} \mathbf{X}_{\text{A,NH}} \\ \mathbf{X}_{\text{B,NH}} \\ \mathbf{X}_{\text{A,HF}} \\ \mathbf{X}_{\text{B,HF}} \\ \mathbf{X}_{\text{A,PT}} \\ \mathbf{X}_{\text{B,PT}} \end{bmatrix}, \tag{6.1}$$

where the rows of $\mathbf{X}_{i,\text{NH}}$ for $i = \{\text{A}, \text{B}\}$ contain the partial averages $\overline{\mathbf{x}}_i$ of every normal hearing subject. The number of rows of $\mathbf{X}_{i,\text{NH}}$ represent the number of normal hearing subjects and is addressed as I_{NH}. Respectively, I_{HF} means the number of patients with high frequency hearing loss and I_{PT} counts the number of patients with pantonal hearing loss. Therefore, the dimension L of the matrix $\mathbf{X}_{\text{data}}$ equals: $L = (I_{\text{NH}} + I_{\text{HF}} + I_{\text{PT}}) \cdot 2$. N represents the length of the partial averages $\overline{\mathbf{x}}_l$ and equals $N = 512$. The WP analysis is applied to $\mathbf{X}_{\text{data}}$ and yields an optimised WP decomposition vector $\mathbf{z}_{\text{opt}}^T$ according to the procedure described in Section 2.4.1.2. The resulting decomposition tree is shown in Figure 6.14.

The figure shows that there are different decomposition depths for the low-pass (LP) as well as for the high-pass (HP) components. For comparison, in a DWT decomposition tree, only the LP component would always be decomposed further until level 9 is reached leaving the HP components undecomposed.

The reason for patching all partial averages $\overline{\mathbf{x}}_l$ into one matrix and not applying the WP analysis to the averaged TEOAE data is that a better entropy minimisation is achieved. Also, one could suggest that separate optimal decompositions for each of the three separation cases are conducted. This is in fact a valid argument and is the objective of the investigations in [21]. However, it was found that this approach shows poorer results for the test data and performs poorer for the control data as well. Therefore, the proposed approach seems to be well suited. For reasons of comparison, the results from [21] will also be shown and discussed.

To confirm the approach of decreasing the entropy with WP compared to the DWT, Table 6.5 shows an overview of the averaged entropies for the transform coefficients $\mathbf{y}_l$. The values in the table correspond to the case that is illustrated in Figure 2.9 where the run of the entropy can be observed for a signal with a length of 512 and an energy content normalised to 1.

Figure 6.15 presents the average WP coefficient energy for the WP decomposition for comparison reasons with the DWT. We see that approximately the same energy distribution is highlighted by the

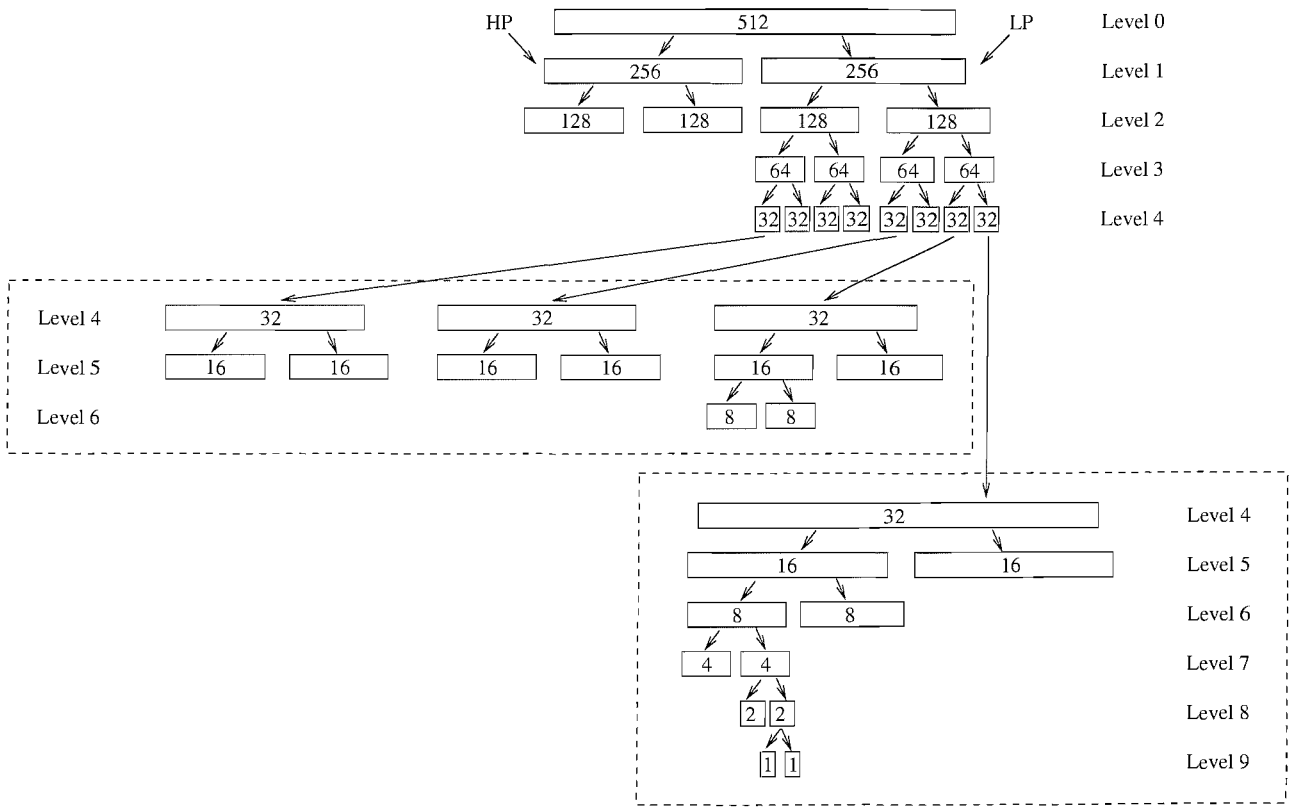

Figure 6.14: WP decomposition structure for Homburg TEOAE data with the low-pass (LP) components on the right and the high-pass (HP) components on the left.

Transformation	entropy	
	Homburg	Heidelberg
DWT	3.643	3.784
WP	3.520	3.654

Table 6.5: Entropy comparison between DWT and WP for the Homburg and Heidelberg data.

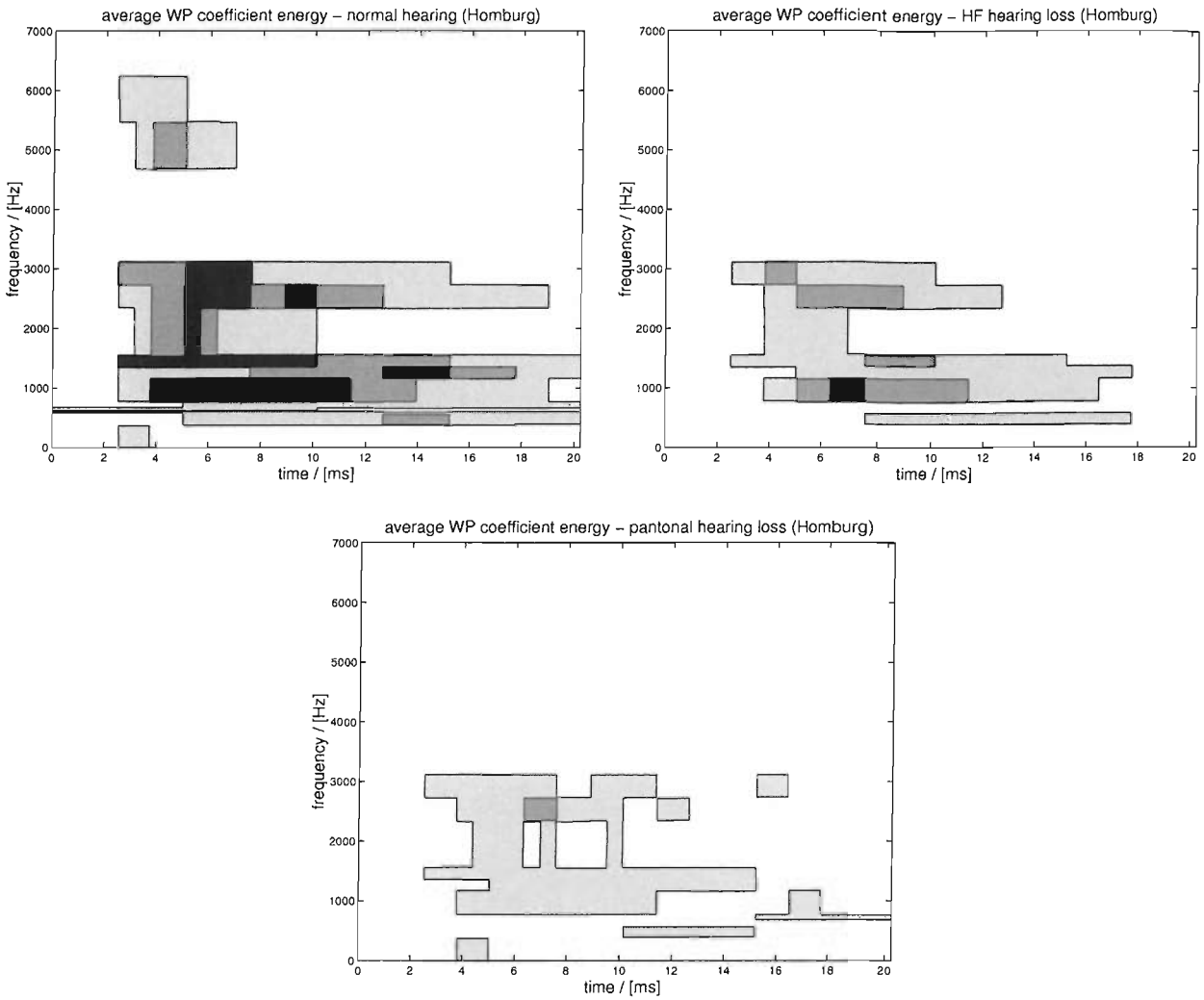

Figure 6.15: Average WP coefficient energy for the Homburg data: (top left) normal hearing (NH), (top right) high frequency hearing loss (HF), and (bottom) pantonal hearing loss (PT).

WP coefficients as for the DWT. This indicates a good parameterisation.

6.3.2 Feature Selection

With the WP decomposition described above, we arrive at a coefficient set C_{opt} as presented in Figure 6.16. Again, the results shown in the figure are optimal in the sense that the best results for the Homburg were attained.

Table 6.6 indicates the respective separability results.

As the WP decomposition is more flexible than the DWT, the search area for significant coefficients is limited to the first order neighbourhood. Second order neighbourhood searches result in incoherent coefficient sets meaning that too much noise is modelled. The resulting coefficient sets confirm this approach as they cover areas which make physiological sense similar to the DWT. Moreover, generalisation is assured as the Heidelberg data shows similar good separability apart from the NH vs HF case, where only a slight improvement to the standard analysis in Table 6.2 can be achieved.

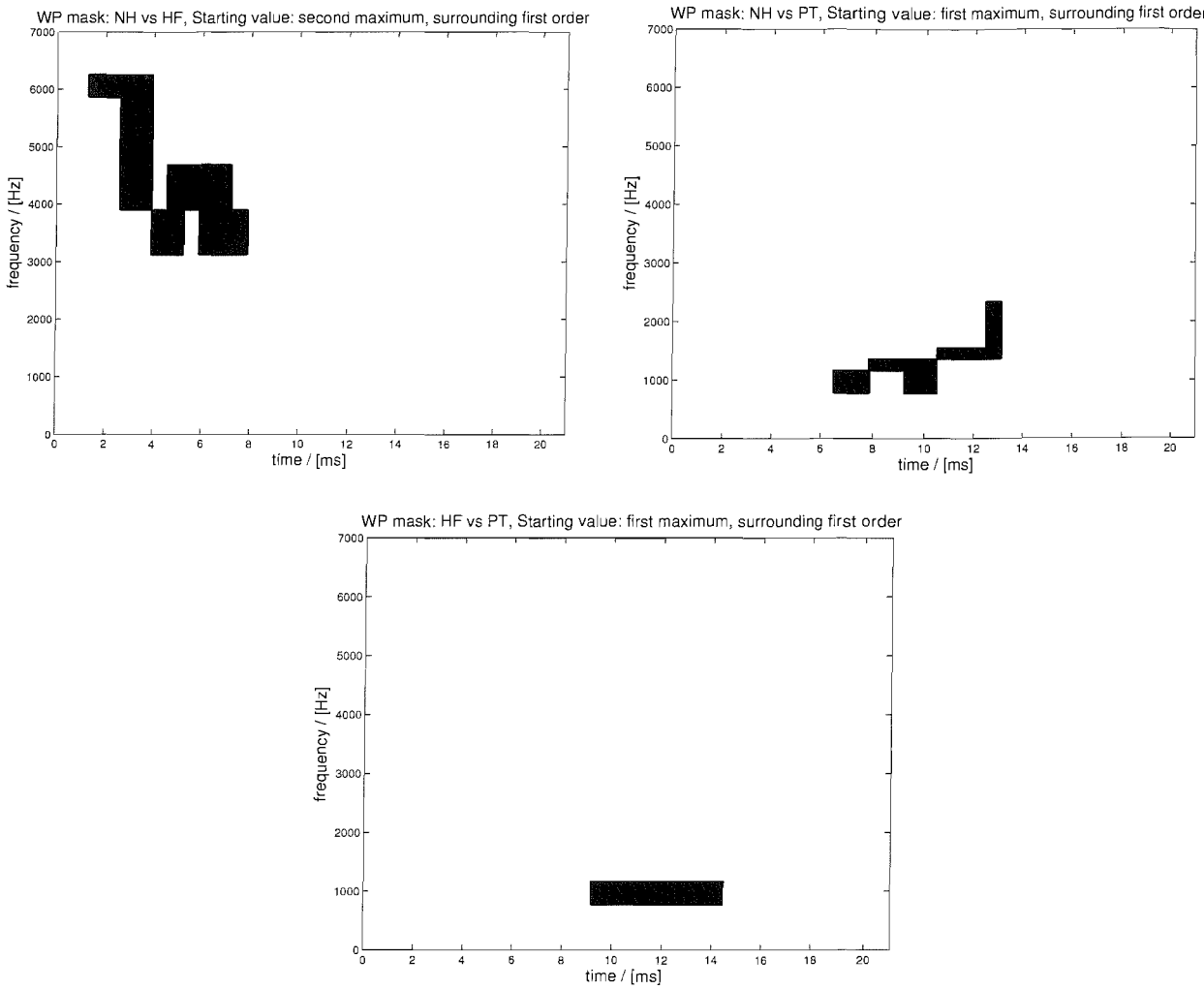

Figure 6.16: Resulting WP coefficients in the TF plane: (top left) NH vs HF yielding 16 coefficients, (top right) NH vs PT yielding 5 coefficients, and (bottom) HF vs PT yielding 4 coefficients.

group distinction	separability	
	Homburg	Heidelberg
NH — HF	0.918	0.799
NH — PT	0.954	0.944
HF — PT	0.843	0.843

Table 6.6: Separability (area under ROC curve) between the 3 hearing ability groups for the Homburg and Heidelberg data for WP.

As mentioned in 6.3.1, we could also calculate WP decomposition for each distinction case separately. Table 6.7 shows the results. It can be observed that the comparison with 6.6 yields poorer

	separability	
group distinction	Homburg	Heidelberg
NH — HF	0.932	0.773
NH — PT	0.990	0.967
HF — PT	0.821	0.837

Table 6.7: Separability (area under ROC curve) between the 3 hearing ability groups for the Homburg and Heidelberg data for a separate WP decomposition for each distinction case.

results for the Homburg data, as well as for the Heidelberg data for the majority of the cases although the Heidelberg data is supposed to be of better quality. Therefore, a uniform WP parameterisation of the data seems to be better suited. The reason for this can be that there is a larger choice when searching for an optimal decomposition among the whole data set as compared to only for data representing one distinction case. Moreover, the stated results in Table 6.7 include also second order neighbourhood search which can be the reason for the adaptation to the Homburg data. Hence, the restriction to the first order neighbourhood search seems reasonable.

Recapitulating it can be said that the WP decomposition shows a slight adaptation to the data used for adjustment. One could suggest using a KLT [28] for parameterisation. However, the findings that a WP decomposition already shows a slight adaptation to the data used for adjustment and parameterises some noise lead to the expectation, that the KLT would not yield good results for the control data for confirming generalisation.

6.3.3 Classification

As for the DWT parameterisation, we firstly show the ROC curves for the WP based SVM classification of the data. Then, we compare and discuss the resulting detection rates for a normal DAGSVM and a DAGSVM with a neutral class.

Again, Figures 6.17–6.19 show that the ROC values for the training are very high for all distinction cases. Moreover, although the ROC values for the feature selection are generally larger for the DWT, the separability values after the classification are larger for the WP than for the DWT especially for the cases NH vs HF and HF vs PT. Again, a one to one comparison between the ROC values after classification and before is faulty because the HF group is split by the DAGSVM decision tree. However, for the WP the separability values are in the same range for the test data as for the standard analysis in Table 6.2 which shows that with our methods a multi-class analysis can have the same significance as a standard comparison for each case against each other. This represents a major result for our studies.

The detection rates for the DAGSVM without neutral class are 68.1% for the NH group, 74.7% for the HF group and 56.4% for the PT group. Compared to the DWT, the HF group detection rate is relatively higher, whereas the other two are smaller. The results for the DAGSVM with neutral class are shown next in Table 6.8. Again, the neutral class can also be observed in Figures 6.18 and 6.19 as

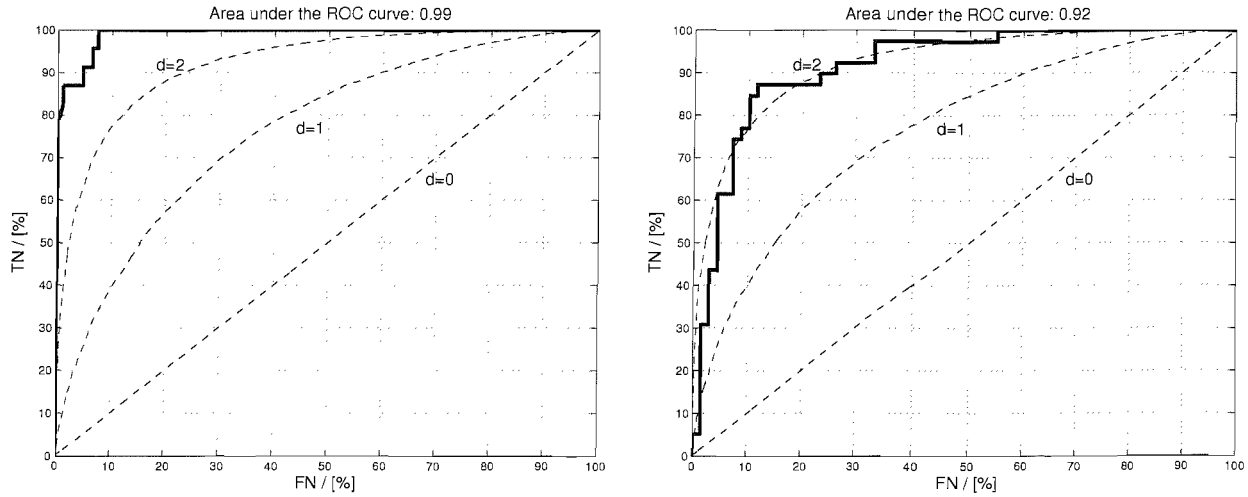

Figure 6.17: ROC curves for the NH vs PT DAGSVM node for (left) training and (right) testing based on a WP parameterisation.

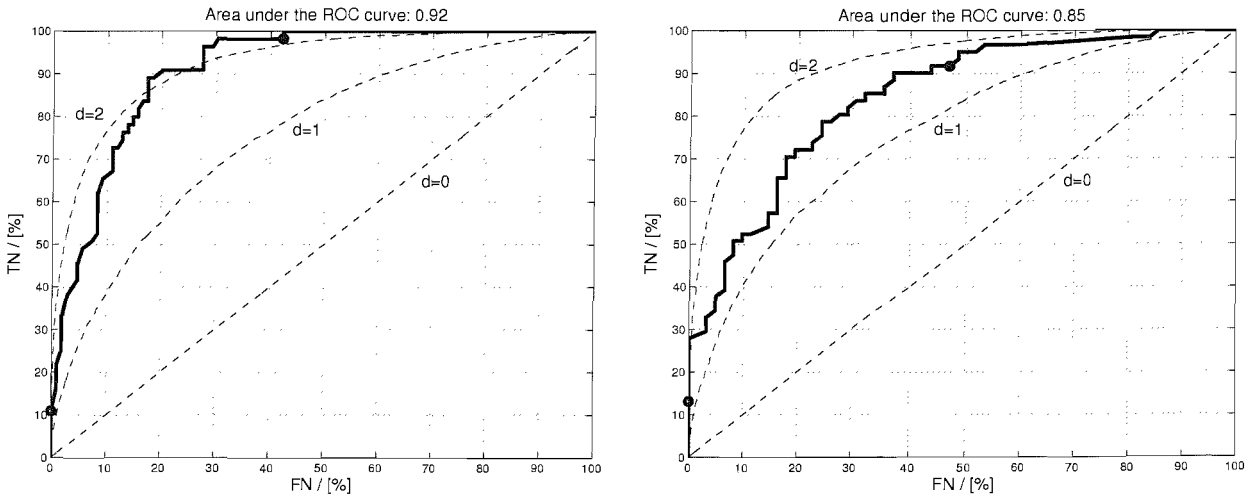

Figure 6.18: ROC curves for the NH vs HF DAGSVM node for (left) training and (right) testing based on a WP parameterisation. The dots indicate the margins for the determined neutral class.

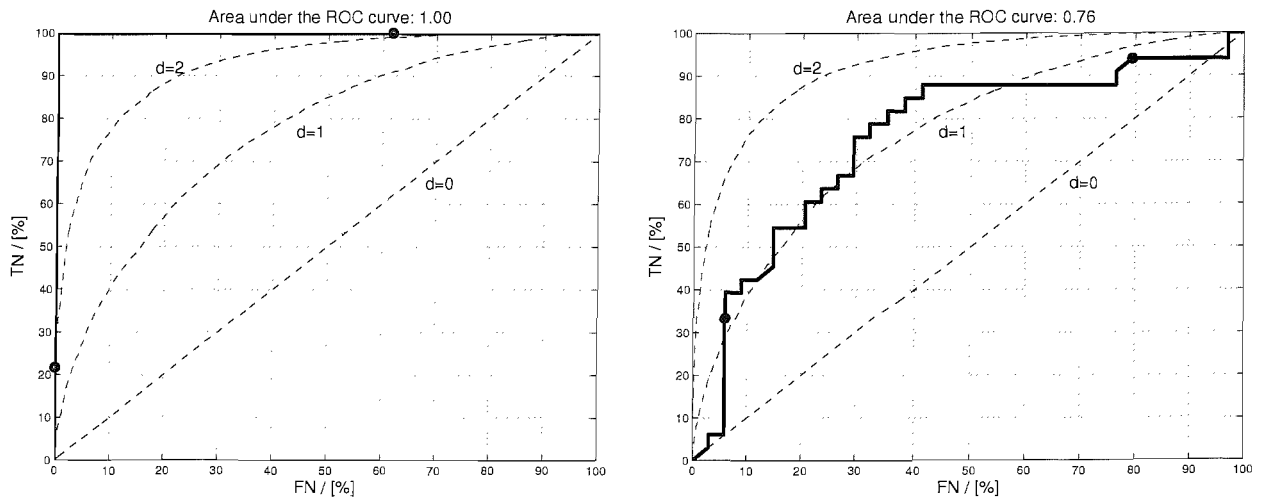

Figure 6.19: ROC curves for the HF vs PT DAGSVM node for (left) training and (right) testing based on a WP parameterisation. The dots indicate the margins for the determined neutral class.

NH false	NH neutral	NH correct	HF false	HF neutral	HF correct	PT false	PT neutral	PT correct
10.1%	42.0%	47.9%	7.4%	76.8%	15.8%	20.5%	51.3%	28.2%

Table 6.8: Detection rates yielded by DAGSVM classification with a neutral class for test data for WP parameterisation.

the area between the dots. We see that NH group can be detected easier than the PT group. The HF group is the most difficult to detect because the more than three quarters of the group are allocated as neutral.

Compared with the DAGSVM without a neutral class, the detectability of the HF group is severely decreased by the neutral class from 74.7% to 15.8%. However, with a rate of 7.4% falsely detected patients, the HF group shows a relatively low error rate. This might also lead to the assumption that for a normal DAGSVM the detection of the HF group is the most untrustworthy.

6.4 GF Analysis of TEOAE

In this section, the results for the feature extraction, feature selection and classification for the GF analysis of TEOAE are shown. Also, we discuss how and why we apply the GF transform to the TEOAE data at the beginning.

6.4.1 Feature Extraction: Transform Adjustment and GF Decomposition

As the GF transform is based on a prototype filter where the channel number, decimation ratio and filter length can be chosen by the user, it can be seen as more flexible than the DWT and WP, where the basis functions are derived from one specific mother wavelet. Hence, the GF transform can be expected to yield better separability results. To find a starting point for the prototype filters to test, the results from the DWT and WP analysis of TEOAE can be used. E.g. for the distinction case NH vs HF, the resulting coefficient sets show a "small and long" characteristic in the TF plane, ranging from 2 to 8 ms and 3 to 6 kHz. Therefore, prototype filters with a relatively small channel number are supposed to parameterise this distinction case better than ones with larger channel numbers, which are more likely to yield significant results for the two other distinction cases.

Only one condition restricts the flexibility of the filter bank design for the GF: the length of $\overline{\mathbf{x}}_i$ holding the partial average needs to be an integer multiple of the time segmentation D. Therefore, when we conduct a GF transform, the first 64 values ($\approx$ 2.5ms) of the data are discarded. Signal information is not lost by this approach as the TEOAE data is expected to be absent for this time period. The reason for this expectation is that during the time period of these first 2.5ms the "click stimulus" is applied as seen in Figure 6.2. The generation of the matrix $\mathbf{H}_{GF}$ follows exactly the same procedure as discussed in Section 2.5.3

Figure 6.20 shows the impulse and magnitude responses of the selected prototypes for the three distinction cases which were chosen after various elementary Gabor functions had been tested as filters.

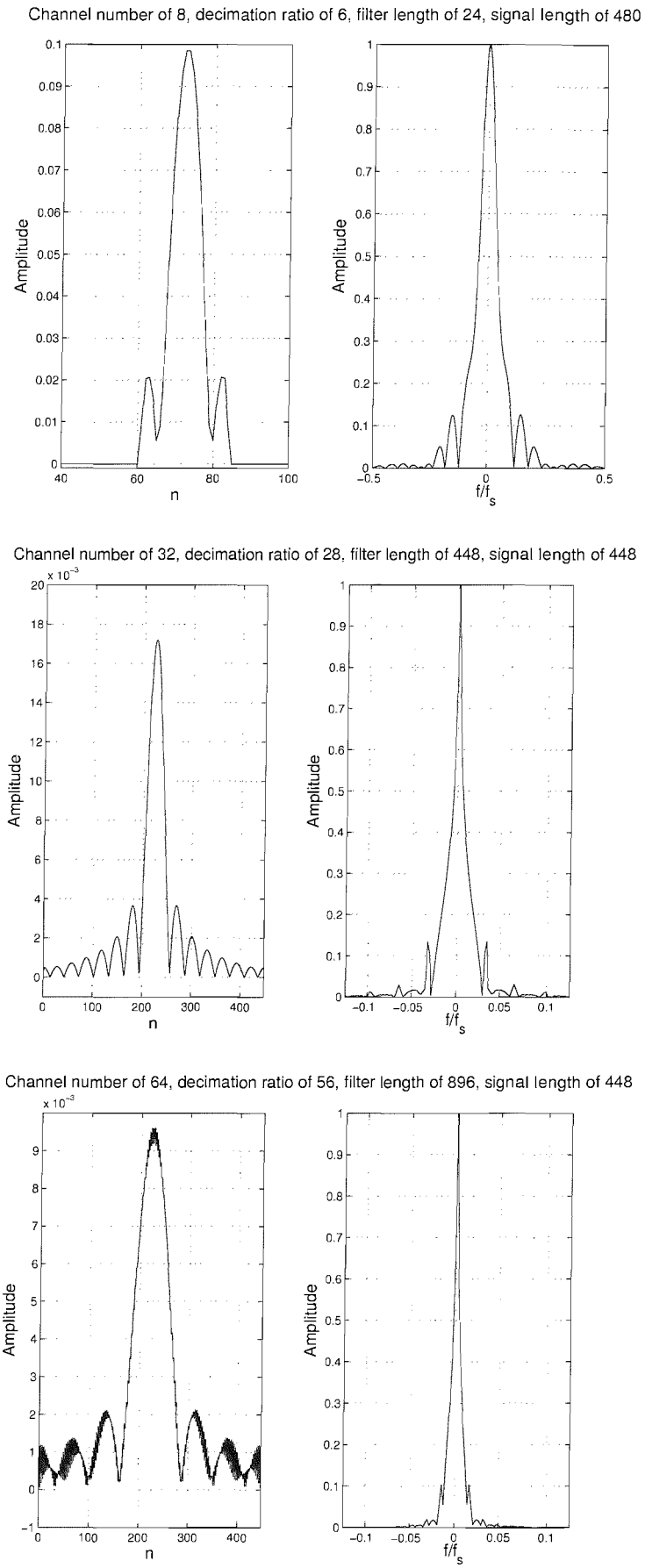

Figure 6.20: (Left) impulse and (right) magnitude response for chosen prototype filters for (top) NH vs HF, (middle) NH vs PT and (bottom) HF vs PT.

The figure represents the absolute value of one row of the respective complex transform matrix $\mathbf{H}_{GF}$ also stating the channel number, decimation ratio, filter length and signal length of the data. It can be seen that e.g. for the case NH vs PT, the signal length is 448 meaning that the first 64 values of the data are skipped to fulfil the above mentioned condition. Moreover, the signal extension can be observed from the impulse response of the chosen filter for the case HF vs PT, which leads to a "smearing" especially at the ends of the discrete filter which is due to the length of the filter with 896. For the case NH vs HF the length of the impulse response is 24 meaning that the signal extension has only an influence when the ends of the time interval are transformed.

According to (2.37) and (2.38), the length of each transformed vector $\mathbf{y}$ is 324 for the case NH vs HF, 272 for the case NH vs PT and 288 for the case HF vs PT. The GF transform of the TEOAE data conducted by the respective transformation matrices lead to an absolute value of the average coefficient energy as shown in Figure 6.21. The comparison of this figure with the corresponding figures for the DWT and WP yields that the GF transform has a more compact TF characteristic than the WP and that in general the TF distribution obtained by the GF transform is more similar to the DWT than to the WP.

6.4.2 Feature Selection

For the GF transform described above, the feature selection method is applied to the absolute values of the transformed vector and yields the following coefficient set C_{opt} as presented in Figure 6.22. Again, the results shown in the figure are optimal in the sense that the best result for the Homburg data was attained. Table 6.9 shows the respective separability results. Similar to the DWT, the extensive simulations for the neighbourhood search revealed a significant adaptation to the Homburg data for the case NH vs PT. Therefore, for this case the search for significant coefficients in the TF plane is restricted to first order only, as this case is the easiest to separate among the three which is indicated by Table 6.2.

The differentiation NH vs HF is obtained by starting with the first maximum and the surrounding of first order. The obtained separability is more balanced for the Homburg and Heidelberg data than the results obtained by DWT or WP. The case NH vs PT also shows a good separability and a reasonable TF distribution of the coefficients. The results are obtained by starting with the first maximum and surrounding of first order. For the case HF vs PT, the results are obtained by starting with the first maximum and surrounding second order with separability values that are in between of the results for the DWT and WP. For the feature selection, the GF show the best overall results. This is likely due to the usage of the flexible prototype filters for creating the transformation matrix $\mathbf{H}_{GF}$. Moreover, the prototypes are chosen based on the results for the DWT and WP analysis and hence, incorporating previous analysis results.

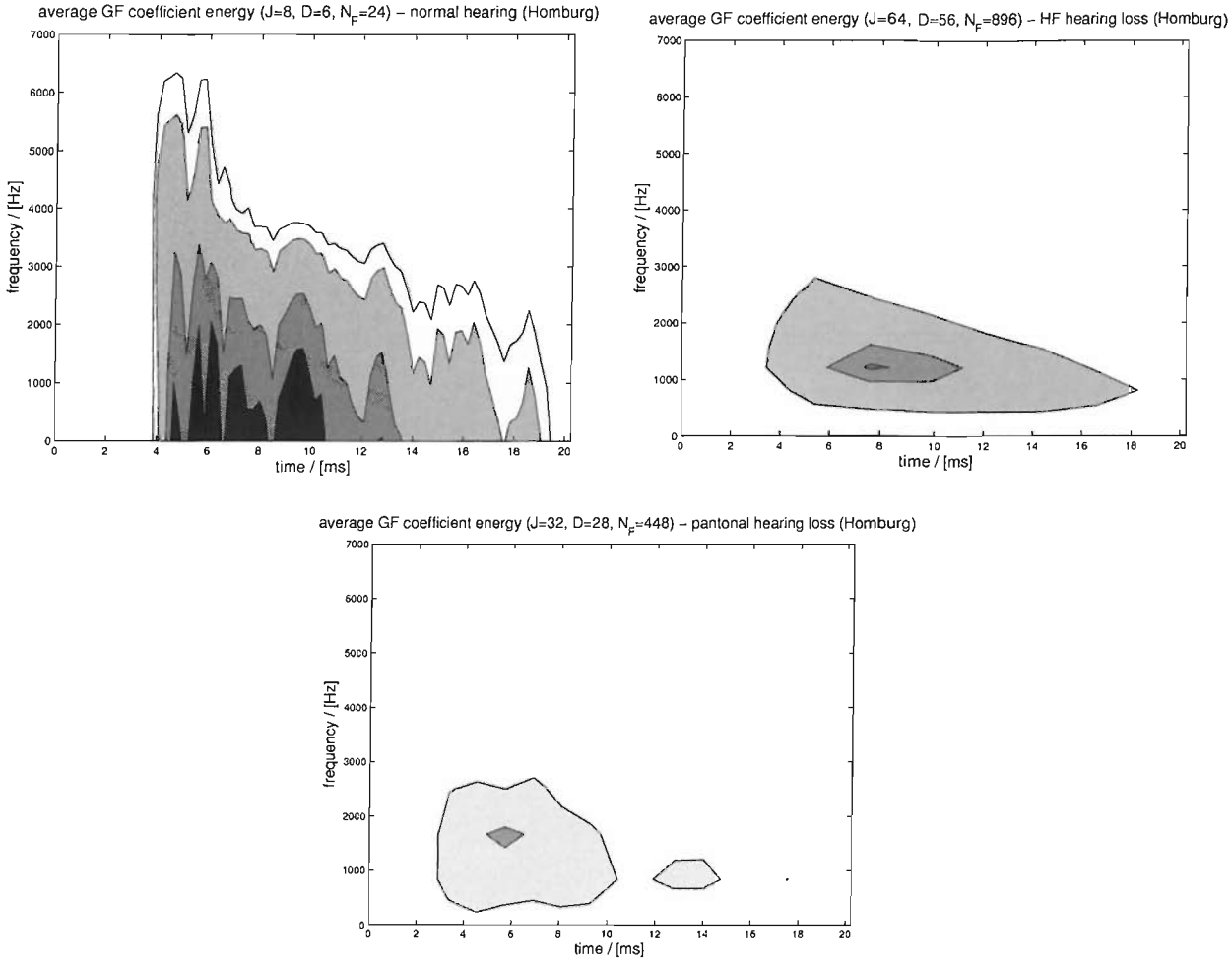

Figure 6.21: Absolute values of the average GF coefficient energy for the Homburg data: (top left) normal hearing (NH), (top right) high frequency hearing loss (HF), and (bottom) pantonal hearing loss (PT).

group distinction	separability	
	Homburg	Heidelberg
NH — HF	0.869	0.829
NH — PT	0.949	0.957
HF — PT	0.840	0.859

Table 6.9: Separability (area under ROC curve) between the 3 hearing ability groups for the Homburg and Heidelberg data for GF.

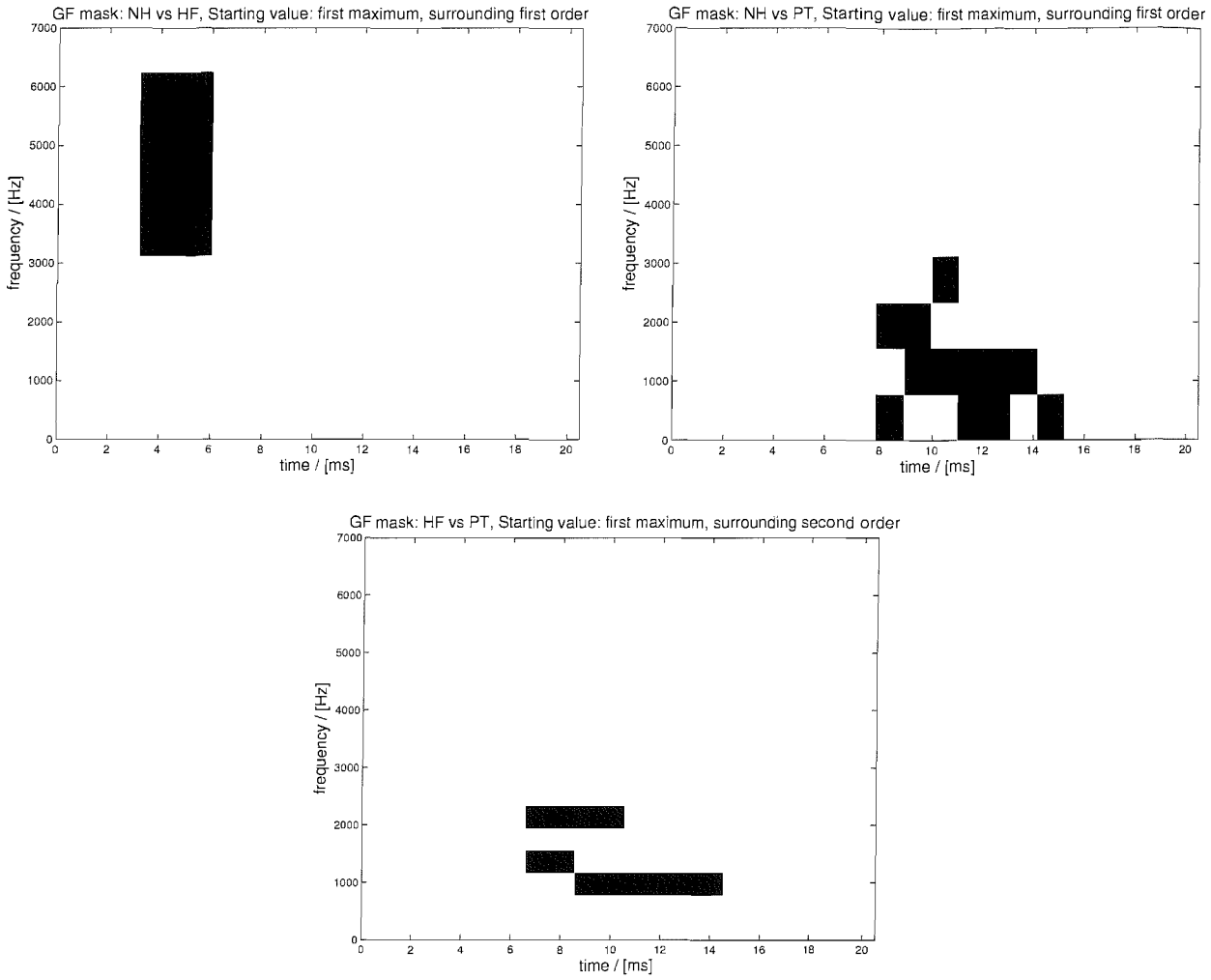

Figure 6.22: Resulting GF coefficients in the TF plane: (top left) NH vs HF yielding 12 coefficients, (top right) NH vs PT yielding 12 coefficients, and (bottom) HF vs PT yielding 6 coefficients.

6.4.3 Classification

Corresponding to the previous parameterisation methods, we give the separability values by the area under the ROC curve for the classification of the GF transformed data first. Then, we show and compare the results for a DAGSVM test and and a DAGSVM test with a neutral class.

Figures 6.23– 6.25 show the ROC curves for the DAGSVM classification. The separability values for the training data are lower compared to the other transforms. Moreover, apart for the case NH vs PT, the separability values for the test data decrease even more than compared to equivalent case for the DWT. This is surprising as the ROC values for the feature selection are the best overall among all transforms.

The detection rates for the test groups for a DAGSVM classification are 91.3% for NH, 63.2% for HF and 53.9% for PT. Table 6.10 show the results for a DAGSVM with a neutral class which can illustratively be observed in Figures 6.24 and 6.25 as the areas between the dots.

For the GF parameterisation the classification of the NH group is very high with 91.3%. However, when applying a neutral class, 45.0% are allocated to that class, but no person is allocated incorrectly

NH false	NH neutral	NH correct	HF false	HF neutral	HF correct	PT false	PT neutral	PT correct
0%	45.0%	55.0%	24.2%	71.6%	4.2%	15.4%	84.6%	0%

Table 6.10: Detection rates yielded by DAGSVM classification with a neutral class for test data for GF parameterisation.

from the NH test group. For the PT group, it is the other way around, no subject is classified correctly, 84.6% are allocated as neutral. Overall, the neutral classes are relatively large for each group.

Having given the classification results for each parameterisation method, we proceed with a comparison and discussion of the results in the following.

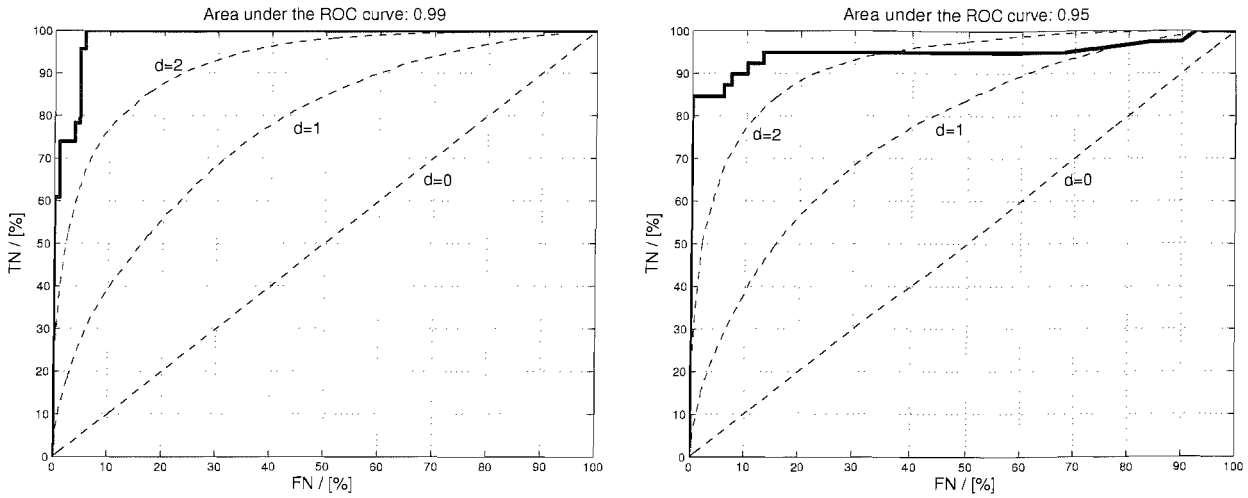

Figure 6.23: ROC curves for the NH vs PT DAGSVM node for (left) training and (right) testing based on a GF parameterisation.

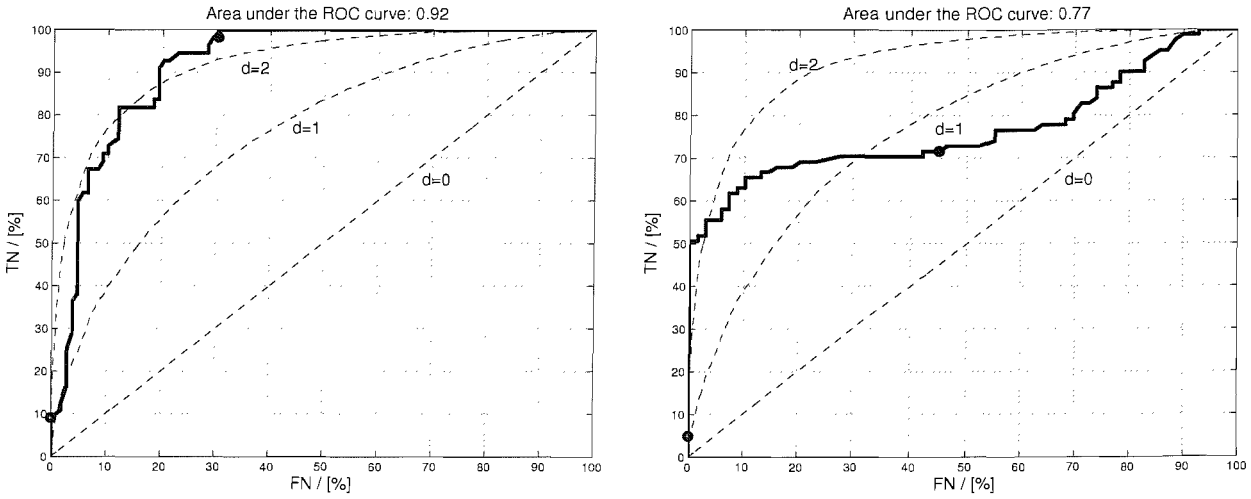

Figure 6.24: ROC curves for the NH vs HF DAGSVM node for (left) training and (right) testing based on a GF parameterisation. The dots indicate the margins for the determined neutral class.

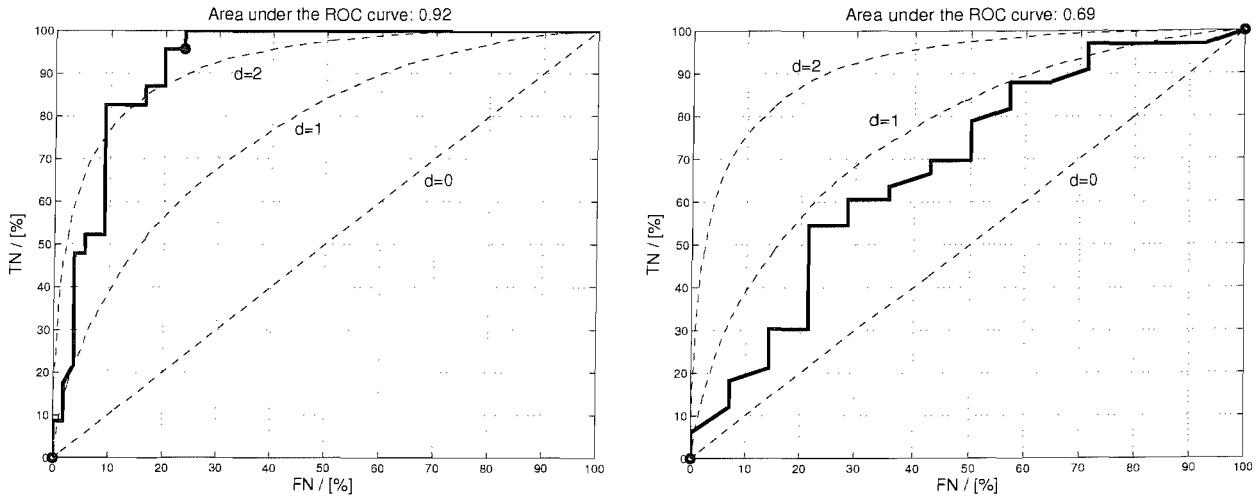

Figure 6.25: ROC curves for the HF vs PT DAGSVM node for (left) training and (right) testing based on a GF parameterisation. The dots indicate the margins for the determined neutral class.

6.5 Comparison of the Results and Discussion

To compare and discuss the results for the different parameterisations, an overview of the detection results for the test data is presented in Table 6.11.

The table shows that the DWT yields the best overall results. The HF can be detected most significantly with the WP. The PT group is the most difficult to determine, just above half of the patients can be allocated correctly for the WP and GF.

The results for the neutral class DAGSVM are shown by Table 6.12.

These results may suggest that the WP outperform the DWT overall as the neutral classes are larger and hence, the error rates are smaller. The GF shows definitely the poorest performance. This could be due to the fact that the GF transform is redundant. When the SVM classification is conducted, the vectors in the input space are not linearly independent. Therefore, the determined SVM classifier cannot divide the data points as well as for orthogonal input parameters which is the case for the DWT and WP. This statement is confirmed when the redundancy is reviewed in more detail. The least redundancy is introduced for the case NH vs PT as the length of the transformed vector $\mathbf{y}$ is the smallest. This case shows separability results in the range of the other transforms in contrast to the other two cases which are tested based on a higher redundancy in the transformed data.

group	Detection rates for test data		
	DWT	WP	GF
NH	79.7%	68.1%	91.3%
HF	63.2%	74.7%	63.2%
PT	69.3%	56.4%	53.9%

Table 6.11: Overview of detection rates yielded by DAGSVM classification for test data.

group	Detection rates for test data		
	DWT	WP	GF
NH false	8.7%	10.1%	0
NH neutral	31.9%	42.0%	45.0%
NH correct	59.4%	47.9%	55%
HF false	24.3%	7.4%	24.2%
HF neutral	57.9%	76.8%	71.6%
HF correct	17.8%	15.8%	4.2%
PT false	28.3%	20.5%	15.4%
PT neutral	25.6%	51.3%	84.6%
PT correct	46.1%	28.2%	0

Table 6.12: Overview of detection rates yielded by DAGSVM classification for test data with neutral class.

Another point to explain the results especially the relatively high values for the neutral class for the HF group is the fact that we used so called unbalanced data, meaning that the classes have unequal class points for training data groups as well as the test data groups, e.g. the NH group is almost three times as large as the PT for the training data. For the l-o-o error estimation that is used to determine the classification network, it is more important to classify the larger class correctly than the smaller class when using unbalanced data. Hence, a larger sensitivity and a small separability can be the consequence. However, this statement also confirms our approach of introducing a neutral class as it is better to being able to say that no decision can be made instead of making many false decisions.

Considering the overall detection results for our system, they may not seem to be encouraging. However, when only considering the the case NH vs PT, the following results are obtained:

- DWT: NH 91.3%, PT 89.7%,
- WP: NH 89.9%, PT 84.6%
- GF: NH 99%, PT 84.6%,

which is well in the range of other studies, e.g. [7]. As stated previously, one major result of our studies is that for the WP the separability values are in the same range for the test data as for the standard analysis in Table 6.2 which shows that with our methods a multi-class analysis can have the same significance as a standard comparison for each class against each other.

In [7], a group of normal hearing is defined by no hearing loss up to 30 dB and a hearing impaired group with a hearing loss over 30 dB. A separation method based on wavelet transforms, ensemble correlation, time window design and mean cross-correlation is introduced. The study concludes with stating the separability values for a hit rate or sensitivity of 90% yielding a value of 65% for only considering the cross-correlation coefficient that is calculated by the measurement equipment. This value is increased by the various methods by approximately 15% to 80% in that study. Compared to our study we achieve slightly better results for the case NH vs PT, which can be seen as equivalent to the case shown in [7] as stated above. One could also argue, that our methods lead to a better separation of hearing loss as our threshold for defining the difference between NH and PT was 20 dB, and the worse the hearing loss gets, the weaker the TEOAE appear and therefore the easier it should be to separate them. On the other hand, as we achieve the lowest value of 65% for the specificity for the case HF vs PT, this shows that it is easier to separate when clear TEOAE are present, which is more likely the case for a threshold of hearing loss of 20 dB than for 30 dB. Moreover, the approach in [7] seems to be narrower, as the threshold for a specific sensitivity value is optimised. Maximising the ROC area is more general, and aims at optimising the specificity for any sensitivity.

The reason why we mainly compare our results with [7] is the similar approach for analysing TEOAE, e.g. applying a wavelet transform for parameterisation. In the following other work on the classification of TEOAE is discussed.

There are studies that deal with the pure detection of TEOAE, meaning classifying if TEOAE are present or not; for example in [82], TEOAE recordings are transformed into a parameter set which represents the input to an artificial neural network. The parameter set contains e.g. the correlation coefficient and 3 parameters based on a subdivision of the correlation coefficient. The results are 99.3%

sensitivity and 81.1% specificity for a training group and 99.4% sensitivity and 87.3% specificity for a test group. Obviously, the applied methods in [82] are significantly different to our approach.

In [4] fast Fourier transformed TEOAE data is studied yielding a sensitivity of 90.2% and specificity of 87.5% when comparing a normal hearing group with a HL group. These results are in the range of the results we obtained with our study.

Recapitulating it can be said that our approach yields separation results that can well compete with other studies so far.

6.6 Summary

We have presented a TF analysis of TEOAE that aims at the detection of frequency specific hearing loss. We have motivated the use of TF methods, and applied a feature selection method to optimise a set of distinctive TF coefficients. This maximisation represents the input to a SVM classifier for the detection. We used two data sets for training and testing. The validity of the results was verified by a test group. Moreover, the results obtained proved to be competitive when they were compared to similar study which also aims at the detection of TEOAE. Therefore, the results appear reasonably robust and encourage frequency specific hearing loss detection via signal processing of TEOAE.

Chapter 7

Conclusions and Future Works

7.1 Conclusions

We have presented digital signal processing methods for the classification of biomedical data. In more detail, linear transformations, namely the DFT, DWT, WP, GF and the KLT were presented with focus on the DWT, WP and GF. For the implementation of the latter transforms, a common matrix notation was introduced. The signal extension necessary when dealing with data on finite support was incorporated in the transform matrix $\mathbf{H}$. The application of these transforms is the parameterisation of data, for which we here considered panic disorder EEG and TEOAE with the aim of extracting their characteristic features. It was found that the DWT, WP and GF are better suited for parameterising or analysing transient biomedical signals, rather than the DFT and KLT, especially when generalisation is taken into account.

The transforms are the basis for identifying the features in the data, which are application-dependent. For their identification, we aim for example to concentrate as much signal energy in as few transform coefficients as possible. Alternatively, we can also aim to find coefficients that provide a good distinction between two or more data sets.

The selected features are the input for a learning machine which is implemented by SVM. Having shown their connection to learning theory by a mathematical formulation of the training procedure, capacity limitation, classification and generalisation, we introduced neutral decisions for multi-class classification for the application of SVM for diagnosis.

For the panic disorder analysis, a novel approach to study the disease was shown by applying the TF transforms to separate data containing the responses to panic causing stimulus from responses to neutral stimulus. Transform coefficients, containing the main features of the data were selected by applying statistical tests resulting in a description of the differences by only two coefficients. The approach was confirmed by comparing it to several other methods including a simple time domain analysis to a SVM study. The results cannot only contribute to the understanding of the disorder but can also be applied, e.g. when investigating the success of a therapy.

The second type of biomedical data studied were TEOAE, where we are trying to maximise the detection between various degrees of frequency specific hearing loss. Based on a SNR-like criterion, an algorithm was developed that searches for a distinctive coefficient set in the TF plane to extract the

features of the data. Then, the data is classified by a multi-class SVM. Moreover, neutral decision were introduced for the SVM classification. The validity of this approach was confirmed by the following main points: Firstly, a test with a control group was conducted. Secondly, the achieved results for the differentiation and detection of hearing loss based on TEOAE were compared with a similar study and revealed a similar or even slightly better performance. Moreover, the general results represent an improvement over previously presented separabilities in [13] due to the improved selection of significant coefficients on which a decision will be based, and by deploying an enhanced criterion, as well as more flexible parameterisation methods. Also, the separability values for the multi-class analysis are in the same range as a standard analysis for each class against each other which shows that with our methods a multi-class analysis can have the same significance as a standard one against one comparison.

7.2 Future Works

Having regarded transforms in the TF domain ranging from fixed to fully data adaptive, it was found that DWT, WP and GF are a reasonable choice to parameterise biomedical data. The applied detection method were SVM. However, neural networks (NN) can be referred to as being the standard classification method. Therefore, it is of interest, what detection results could be achieved with NN and how they compare with the SVM results. Also, to make the detection more robust, a sequential statistical tests could be applied. E.g. for the TEOAE analysis, if the data is not averaged and each of the 520 measurements are analysed, a statistical test after the classification could yield a decision after a certain number of measurements analysed making the the analysis of all measurements dispensable.

Concerning the SVM classification of TEOAE, the neutral class method could be explored in more detail. The following questions seem to be the most interesting to answer:

- What results are obtained when the neutral class for the SVM is also applied to the distinction case NH vs PT?
- How can reasonable thresholds be determined for the neutral class? What sensible approach can be used for that?
- What are the results for balanced data sets?

The main reason for not using a neutral class for the NH vs PT classification was that the coefficient set used for this distinction was specially adapted to these two groups and the results for the HF class are therefore unpredictable. So, the determination of a coefficient set for all three classes and a classification with the same thresholds for the neutral classes may be interesting to investigate which leads directly to the question how to select a threshold for the neutral class and with what approach. As we explained the threshold for the neutral class for our studies was as large as possible, leaving no errors for the training data.

Also, our data sets were unbalanced, the NH group was almost three times as big as the PT group. Therefore, it can be expected that the detection rates for balanced data would show a more even characteristic meaning that the error rates are of similar size. For our studies the detection rates for the NH group proved to be the largest generally speaking.

Furthermore, the analysis of multichannel data, by which spatial information is obtained, for example by recording more than one electrode during an EEG measurement, could lead to another interesting area for the application of SVM based on TF transforms.

Bibliography

[1] R. Northrop, *Noninvasive Instrumentation and Measurement in Medical Diagnosis.* Interpharm/CRC, 2001.

[2] M. Robinette and T. Glattke, *Otoacoustic Emissions: Clinical Applications.* Thieme Medical Pub, 2. ed., 2001.

[3] S. Windmann, Z. Sakhavat, and M. Kutas, "Electrophysiological Evidence Reveals Affective Evaluation Deficits Early in Stimulus Processing in Patients With Panic Disorder," *Journal of Abnormal Psychology*, vol. 111, pp. 357–369, November 2002.

[4] S. Hatzopoulos, M. Mazzoli, and A. Martini, "Identification of Hearing Loss Using TEOAE Descriptors: Theoretical Foundations and Preliminary Results," *Audiology*, vol. 34, pp. 248–259, 1995.

[5] U. Hoppe, S. Weiss, R. W. Stewart, and U. Eysholdt, "An Automatic Sequential Recognition Method for Late Auditory Evoked Responses," *IEEE Transactions on Biomedical Engineering*, vol. 48, pp. 154–164, February 2001.

[6] E. Stürzebecher, M. Cebulla, and K. Wernecke, "Objective detection of transiently evoked otoacoustic emissions," *Scandinavian Audiology*, vol. 29, pp. 78–88, 2001.

[7] A. Janusauskas, L. Sornmo, O. Svensson, and B. Engdahl, "Detection of Transient-Evoked Otoacoustic Emissions and the Design of Time Windows," *IEEE Transactions on Biomedical Engineering*, vol. 49, pp. 132–139, February 2002.

[8] H. Liu, T. Zhang, and F. Yang, "A Multistage, Multimethod Approach for Automatic Detection and Classification of Epileptiform EEG," *IEEE Transactions on Biomedical Engineering*, vol. 49, pp. 1557–1566, December 2002.

[9] V. Vapnik, *Statistical Learning Theory.* New York: Wiley: Cambridge University Press, 1998.

[10] S. Minfen, S. Lisha, and F. Chan, "A novel method for extracting time-varying rhythms of electroencephalography via wavelet packet analysis," in *Proceedings of First International Conference on Advances in Medical Signal and Information Processing*, IEE Conf. Publ. No. 476, pp. 73-78, September 2000.

[11] S. Blanco, C. Attellis, S. Isaacson, O. Rosso, and R. Sirne, "Time-frequency analysis of electroencephalogram series. II. Gabor and wavelet transforms," *Physical Review E*, vol. 54, pp. 6661–6672, December 1996.

[12] F. Brauer, B. Dick, G. Stroink, J. Connolly, P. McGrath, and G. Finley, "KLT analysis of brain potential maps during pain," in *Proceedings of the 22nd Annual International Conference of the IEEE Engineering in Medicine and Biology Society*, vol. 4, pp. 2755–2756, July 2000.

[13] S. Weiss, U. Hoppe, M. Schabert, and U. Eysholdt, "Wavelet Analysis of Transient Evoked Otoacoustic Emissions for Differential Diagnosis of Cochlear Hearing Loss," in *Asilomar Conference on Signals, Systems, and Computers*, (Monterey, CA), November 2001.

[14] P. Pauli, G. Dengler, G. Wiedemann, P. Montoya, H. Flor, N. Birbaumer, and G. Buchkremer, "Behavioural and Neurophysiological Evidence for Altered Processing of Anxiety-Related Words in Panic Disorder," *Journal of Abnormal Psychology*, vol. 106, no. 2, pp. 213–220, 1997.

[15] I. Kitsas, L. Hadjileontiadis, and S. Panas, "Short-term analysis of heart-rate variability using wavelet packets: an efficient detector of sleep apnea episodes," in *Proceedings of the Second Joint EMBS/BMES Conference*, vol. 1, pp. 88–89, 2002.

[16] H. Dietl, S. Weiss, and P. Pauli, "Time-Frequency Transform Based Panic Disorder Classification," in *Proceedings of the 2nd IEEE EMBSS Postgraduate Conference*, IEE, Birmingham, July 2003.

[17] H. Dietl and S. Weiss, "Categorisation of Panic Disorder by Time-Frequency Methods," in *Asilomar Conference on Signals, Systems, and Computers*, (Monterey, CA), November 2003.

[18] H. Dietl and S. Weiss, "Parameterisation Comparison for the Detection of Panic Disorder Using Time-Frequency Transforms and Support Vector Machines," in *Proceedings of the 2nd International Conference on Advances in Medical Signal and Information Processing, MEDSIP*, IEEE, Malta, September 2004.

[19] H. Dietl, S. Weiss, and U. Hoppe, "Comparison of Transformation Methods to Determine Frequency Specific Cochlear Hearing Loss Based on TEOAE," in *Proceedings of IEE Colloquium on Medical Applications of Signal Processing*, vol. 02/110, pp. 16/1–16/6, IEE, London, October 2002.

[20] H. Dietl and S. Weiss, "Difference evaluation method for differential diagnosis of frequency specific hearing loss," in *Proceedings of the IPEM meeting on Signal Processing Applications in Clinical Neurophysiology*, IPEM, York, February 2004.

[21] H. Dietl and S. Weiss, "Parameterisation of transient evoked otoacoustic emissions," in *Proceedings of the BIOSIGNAL 2004 International EURASIP Conference*, Brno, June 2004.

[22] H. Dietl and S. Weiss, "Cochlear Hearing Loss Detection System Based on Transient Evoked Otoacoustic Emissions," in *Proceedings of the 3rd IEEE EMBSS Postgraduate Conference*, IEE, Southampton, August 2004.

[23] H. Dietl and S. Weiss, "Detection of Cochlear Hearing Loss Applying Wavelet Packets and Support Vector Machines," in *Asilomar Conference on Signals, Systems, and Computers*, (Monterey, CA), November 2004.

[24] H. Dietl and S. Weiss, "A Novel Approach to Detect Cochlear Hearing Loss," in *Proceedings of the 6th International Conference on Mathematics in Signal Processing*, Royal Agricultural College, Cirencester, December 2004.

[25] G. Strang, *Linear Algebra and Its Applications.* New York: Academic Press, 2nd ed., 1980.

[26] G. Strang and T. Nguyen, *Wavelets and Filter Banks.* Wellesley, MA: Wellesley–Cambridge Press, 1996.

[27] H. Stöcker, *Taschenbuch mathematischer Formeln und moderner Verfahren.* Frankfurt am Main, Thun, Germany: Verlag Harri Deutsch, 4 ed., 1999.

[28] G. H. Golub and C. F. Van Loan, *Matrix Computations.* Baltimore, Maryland: John Hopkins University Press, 3rd ed., 1996.

[29] M. Wall, A. Rechtsteiner, and L. Rocha, "Singular value decomposition and principal component analysis," in *A Practical Approach to Microarray Data Analysis* (D. Berrar, W. Dubitzky, and M. Granzow, eds.), pp. 91–109, Kluwer: Norwell, MA, 2003.

[30] S. G. Mallat, "A Theory for Multiresolution Signal Decomposition: The Wavelet Representation," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 11, pp. 674–692, July 1989.

[31] I. Daubechies, *Ten Lectures on Wavelets.* Philadelphia: SIAM, 1992.

[32] S.-C. Pei and M.-H. Yeh, "An Introduction to Discrete Finite Frames," *IEEE Signal Processing Magazine*, vol. 14, pp. 84–96, November 1997.

[33] A. Graps, "An Introduction to Wavelets," *IEEE Computational Science and Engineering*, vol. 2, pp. 50–61, Summer 1995.

[34] S. Weiß, "Rekonstruktion Akustisch Evozierter Potentiale," tech. rep., Univ.-HNO-Klinik Erlangen / Lehrstuhl für Technische Elektronik, Universität Erlangen-Nürnberg, Germany, Dec. 1994.

[35] N. J. Fliege, *Multiraten-Signalverarbeitung.* Stuttgart: B.G.Teubner, 1993.

[36] R. E. Crochiere and L. R. Rabiner, *Multirate Digital Signal Processing.* Englewood Cliffs, NJ: Prentice Hall, 1983.

[37] M. Vetterli and C. Herley, "Wavelets and Filter Banks: Theory and Design," *IEEE Transactions on Signal Processing*, vol. Vol.40, pp. 2207–2232, September 1992.

[38] A. Jensen and A. la Cour Harbo, *Ripples in Mathematics.* Berlin, Heidelberg, New York: Springer-Verlag, 2001.

[39] X. Miao and W. Moon, "Application of Wavelet transform in seismic data analysis," *Geosciences Journal*, vol. 3, no. 3, pp. 171–179, 1999.

[40] C. Grimm, "Vorlesung systemtheorie," tech. rep., Technische Informatik, Johann Wolfgang Goethe-Universität, October 2001.

[41] P. Wolfe, S. Godsill, and M. Dörfler, "Multi-Gabor Dictionaries for Audio Time-Frequency Analysis," *IEEE Workshop on Applications of Signal Processing to Audio and Acoustics*, pp. 43–46, October 2001.

[42] M. Zibulski and Y. Zeevi, "Discrete Multiwindow Gabor-Type Transforms," *IEEE Transactions on Signal Processing*, vol. 45, pp. 1428–1442, 1997.

[43] M. Kjeldsen and R. Arndt, "Joint Time Frequency Analysis Techniques: A study of Transitional Dynamics in Sheet/Cloud Cavitation," *Fourth International Symposium on Cavitation*, June 2001.

[44] S. Weiss, L. Lampe, and R. W. Stewart, "Efficient Subband Adaptive Filtering with Oversampled GDFT Filter Banks," in *Digest IEE Colloquium on Adaptive Signal Processing for Mobile Communication Systems*, (London, England), pp. 4.1–4.9, October 1997.

[45] M. Harteneck, S. Weiss, and R. W. Stewart, "Design of Near Perfect Reconstruction Oversampled Filter Banks for Subband Adaptive Filters," *IEEE Transactions on Circuits & Systems II*, vol. 46, pp. 1081–1086, August 1999.

[46] H. G. Feichtinger and T. Strohmer, eds., *Gabor Analysis and Algorithms — Theory and Applications*. Boston, MA: Birkhäuser, October 1997.

[47] S. Weiss and R. W. Stewart, "Fast Implementation of Oversampled Modulated Filter Banks," in *3rd European DSP Education and Research Conference*, (Paris), September 2000.

[48] S. Weiss and R. W. Stewart, "Fast Implementation of Oversampled Modulated Filter Banks," *IEE Electronics Letters*, vol. 36, pp. 1502–1503, August 2000.

[49] J. A. Hanley and B. J. McNeil, "The Meaning and Use of the Area under a Receiver Operating Characteristic (ROC) Curve," *Radiology*, vol. 143, pp. 26–36, 1982.

[50] K. Chu, "An introduction to sensitivity, specificity, predictive values and likelihood ratios," *Emergency Medicine*, vol. 11, pp. 175–181, 1999.

[51] P. Armitage, G. Berry, and J. Matthews, *Statistical Methods in Medical Research*. Oxford: Blackwell Science, fourth ed., 2002.

[52] Otodynamics Ltd., Hatfield, Herfortshire, UK, *ILO OAE Instrument User Manual*, 5a ed., October 1997.

[53] N. Cristianini and J. Shawe-Taylor, *An introduction to Support Vector Machines*. Cambridge, United Kingdom: Cambridge University Press, 2000.

[54] C. Burges, "A tutorial on support vector machines for pattern recognition," *Data Mining and Knowledge Discovery*, vol. 2(2), pp. 121–167, 1998.

[55] K. Műller, S. Mika, G. Ratsch, K. Tsuda, and B. Schőlkopf, "An Introduction to Kernel-Based Learning Algorithms," *IEEE Transactions on Neural Networks*, vol. 12, pp. 181–201, March 2001.

[56] C. Bahlmann, B. Haasdonk, and H. Burkhardt, "On-line Handwriting Recognition with Support Vector Machines – A Kernel Approach," in *Proceedings of the 8th Int. Workshop on Frontiers in Handwriting Recognition (IWFHR)*, pp. 49–54, 2002.

[57] T. Joachims, "Estimating the Generalization Performance of a SVM Efficiently," in *Proceedings of the International Conference on Machine Learning (ICML)*, Morgan Kaufman 2000.

[58] M. Hearst, S. Dumais, E. Osman, J. Platt, and B. Schölkopf, "Support vector machines," *IEEE Intelligent Systems*, vol. 13, issue 4, pp. 18–28, July-Aug. 1998.

[59] G. Cawley, "MATLAB support vector machine toolbox (v0.55β), [`http://theoval.sys.uea.ac.uk/~gcc/svm/toolbox`]." University of East Anglia, School of Information Systems, Norwich, Norfolk, U.K. NR4 7TJ, 2000.

[60] C. C. Chang and C. J. Lin, "Training ν-Support Vector Classifiers: Theory and Algorithms," *Neural Computation*, vol. 13(9), pp. 2119–2147, September 2001.

[61] J. Weston and C. Watkins, "Support vector machines for multi-class pattern recognition," in *Proceedings of the Seventh European Symposium On Artificial Neural Networks (ESANN)*, 1999.

[62] J. Platt, N. Cristianini, and J. Shawe-Taylor, "Large Margin DAGs for Multiclass Classification," in *Proceedings of Advances in Neural Information Processing Systems, NIPS'99*, pp. 547–553, MIT Press 2000.

[63] M. Teplan, "Fundamentals of EEG Measurement," *Measurement Science Review*, vol. 2, sec. 2, pp. 1–11, 2002.

[64] D. M. Clark, "A cognitive approach to panic," *Behaviour Research and Therapy*, vol. 24, pp. 461–470, 1986.

[65] E. A. Kostandov and Y. L. Arzumanov, "Average cortical evoked potentials to recognised and nonrecognised verbal stimuli," *Acta Neurobiologica Experimentalis*, vol. 37, pp. 311–324, 1977.

[66] E. Naumann, D. Bartussek, O. Diedrich, and M. Laufer, "Assessing cognitive and affective information processing functions of the brain by means of the late positive complex of the event-related potential," *Journal of Psychophysiology*, vol. 6, pp. 285–298, 1992.

[67] S. G. Mallat, "Multiresolution Approximations and Wavelet Orthonormal Bases of $L^2(R)$," *Transactions of the American Mathematical Society*, vol. 315, pp. 69–87, September 1989.

[68] D. T. Kemp, "Stimulated Acoustic Emissions from within the Human Auditory System," *Journal of the Acoustic Society of America*, vol. 64, pp. 1386–1391, 1978.

[69] D. T. Kemp, "Otoacoustic Emissions, Travelling Waves and Cochlear Mechanism," *Ear & Hearing*, vol. 22, pp. 95–104, 1986.

[70] E. de Boer, "On Equivalence of Locally Active Models of the Cochlea," *Journal of the Acoustic Society of America*, vol. 98, pp. 1400–1409, 1995.

[71] T. Fukazawa, "A Model of Cochlear Micromechanics," *Hearing Research*, vol. 113, pp. 182–190, 1997.

[72] S. Hatzopoulos, J. Cheng, A. Grzanka, and A. Martini, "Time-Frequency Analyses of TEOAE Recordings from Normals and SNHL Patients," *Audiology*, vol. 39, pp. 1–12, 2000.

[73] P. Ravazzani, G. Tognola, F. Grandori, and J. Ruohonen, "Two-Dimensional Filter to Facilitate Detection of Transient-Evoked Otoacoustic Emissions," *IEEE Transactions on Biomedical Engineering*, vol. 45, pp. 1089–1096, 1998.

[74] M. Whitehead, A. Jimenez, B. Stagner, M. McCoy, B. Lonsbury-Martin, and G. Martin, "Time-Windowing of Click-Evoked Otoacoustic Emissions to Increase Signal-to-Noise Ratio," *Ear & Hearing*, vol. 16, pp. 599–611, 1995.

[75] E. G. Pasanen, J. D. Travis, and R. J. Thornhill, "Wavelet-Type Analysis of Transient-Evoked Otoacoustic Emissions," *Biomedical Sciences Instrumentation*, vol. 30, pp. 75–80, 1994.

[76] G. Tognola, F. Grandori, and P. Ravazzani, "Time-frequency distributions of click-evoked otoacoustic emissions," *Hearing Research*, vol. 106, pp. 112–122, 1997.

[77] J. Cheng, "Time-Frequency Analysis of Transient Evoked Otoacoustic Emissions Via Smoothed Pseudo Wigner-Ville Distribution," *Scandinavian Audiology*, vol. 24, pp. 91–96, 1995.

[78] K. J. Blinowska, P. J. Durka, A. Skierski, F. Grandori, and G. Tognola, "High Resolution Time-Frequency Analysis of Otoacoustic Emissions," *Technology and Health Care*, vol. 5, pp. 407–418, 1997.

[79] N. Nicolaou and S. Nasuto, "Temporal Independent Component Analysis for Automatic Artefact Removal From EEG," in *Proceedings of the 2nd International Conference on Advances in Medical Signal and Information Processing, MEDSIP*, pp. 153–160, IEEE, Malta, September 2004.

[80] S. Ranganatha, C. Guan, M. Thulasidas, W. Xu, X. Zhu, and J. Wu, "Comparison of Artefact Removal Methods on Their Effect on Motor Task and Imagery Classification in a Brain-Computer Interface," in *Proceedings of the 2nd International Conference on Advances in Medical Signal and Information Processing, MEDSIP*, pp. 117–123, IEEE, Malta, September 2004.

[81] E. Bartnik, K. Blinowska, and P. Durka, "Single Evoked Potential Reconstruction by Means of Wavelet Transform," *Biological Cybernetics*, vol. 67, pp. 175–181, February 1992.

[82] G. Buller and M. Lutman, "Automatic classification of transiently evoked otoacoustic emissions using an artificial neural network," *British Journal of Audiology*, vol. 32, pp. 235–247, 1998.

Glossary

ANOVA	analysis of variances
CWT	continuous wavelet transform
DAGSVM	directed acyclic graph support vector machine
DFT	discrete Fourier transform
DPOAE	distortion product OAE
DWT	discrete wavelet transform
EEG	electroencephalogram
EOAE	evoked OAE
ERP	event related brain potentials
FFT	fast Fourier transform
FIR	finite impulse response
FN	false negative rate
FP	false positive rate, false alarm rate
GDFT	generalised DFT
GF	Gabor frames
HF	high frequency hearing loss
HL	hearing loss
KKT	Karush-Kuhn-Tucker
KLT	Karhunen-Loeve transform
l-o-o	leave one out
MRA	multi-resolution algorithm
NH	normal hearing
NN	neural networks
NPV	negative predictive value
OAE	otoacoustic emissions

P300	characteristic transient waveform in ERP
PPV	positive predictive value
PT	pantonal hearing loss
QMF	quadrature mirror filter
ROC	receiver operating characteristic
SFOAE	stimulus-frequency OAE
SMO	sequential minimal optimisation
SNR	signal-to-noise ratio
SOAE	spontaneous OAE
SPL	sound pressure level
SRM	structural risk minimisation
STFT	short time Fourier transform
SVD	singular value decomposition
SVM	support vector machines
TEOAE	transient evoked OAE
TF	time-frequency
TN	true negative rate, specificity
TP	true positive rate, sensitivity, hit rate
WP	wavelet packets

List of Symbols

General Notations

$\mathbf{h}$	vector quantity
$\overline{\mathbf{h}}$	mean vector quantity
$\tilde{\mathbf{h}}$	symmetrically extended vector quantity
$\mathbf{h}_{norm}$	norm of a vector quantity
$\|\|\mathbf{h}\|\|$	Euclidean vector norm
$\mathbf{H}$	matrix quantity
$h(t)$	function of a continuous variable t
$h[n]$	function of a discrete variable n

Relations and Operators

$(\cdot)^{-1}$	inverse
$(\cdot)^*$	conjugated complex
$(\cdot)^T$	transpose
$(\cdot)^H$	Hermitian (conjugate transpose)
$(\cdot)^\dagger$	pseudo inverse
$(\cdot)^F$	"Flip" operator
$< (\cdot), (\cdot) >$	inner product of discrete functions
null$\{\mathbf{H}\}$	null-space of $\mathbf{H}$: $\{\mathbf{x} : \mathbf{Hx} = 0\}$
rank$\{\mathbf{H}\}$	rank of $\mathbf{H}$ (number of linearly independent rows)
C_{opt}	optimal coefficient set containing noisy separation signal information
$C_{\text{opt}}^{\perp}$	coefficient set excluding C_{opt} containing noise only
$\mathcal{ENT}$	procedure to determine WP coefficients by minimising the entropy
$\mathcal{E}\{\cdot\}$	expectation operator
G_g	discrete Gabor transform
jc	complex operator
ln	natural logarithm
SNR_{EN}	signal-to-noise ratio for energies
$STFT_g$	short time Fourier transform
$Var(\cdot)$	variance of a distribution
W_Ψ	wavelet transform
$\mathbf{z}_{\text{opt}}^T$	vector containing the optimised WP decomposition based on the introduced entropy method

Symbols and Variables

A	frame bound
a	continuous scaling parameter
B	frame bound
b	continuous translation parameter
C	parameter which is multiplied with the sum of the slack variables for limitation of the error for SVM classification
c	general constant
c_{data}	constant containing signal information
$c_{j,k}$	Gabor transformed time-frequency coefficients
D	decimation ratio
d	distance of means measured in standard deviations
d_p	order of polynomial kernel for SVM
E	energy of a discrete function
F	feature space for SVM
F_r	variance ratio
F^*_{SP}	modified variance ratio
F_v	F-value for F-test
f_a, f_s	sampling frequency
$g[n]$	elementary prototype for Gabor transform
H	Hilbert space
h_{VC}	Vapnik-Chervonenkis dimension
i	general index
J	scaling or channel number for a time-frequency transform
j	discrete scaling or channel index
j_T	transform selection index
K	length of time-frequency transform vector
K_{FB}	discrete length of frequency band for a Gabor transform
$K_{SVM}(x)$	kernel function for SVM
k	discrete translation index, transformed domain index
k_d	rank of rank-deficient matrix
k_t	truncation limit of a SVD decomposed matrix
L	number of measurements
$L(x)$	Lagrange function
$L'(x)$	Lagrange dual function

l	index for number of measurements
N_F	length of a discrete filter
P	significance level
P_p	distribution probability
$\mathbf{p}_\alpha$	vector holding parameters of a learning machine
$p[n]$	prototype FIR filter for Gabor transform
$R(\mathbf{p}_\alpha)$	actual risk, expected error for a trained learning machine
R_{emp}	empirical risk, error on training data for a learning machine
s_i	indexed noise: random variable with zero mean and variance σ^2
$\text{sign}(x)$	sign function yielding sign of variable
$t_j[n]$	modulation function for discrete Gabor transform
t_v	t-value for t-test
ut	ut-value for ut-test
w	weights for SVM
γ	small constant
γ_R	radius of radial basis function kernel
δ	Dirac impulse
δ_{VC}	probability for VC-confidence
ϵ	entropy according to Shannon
ξ	slack variable representing errors for SVM classification
ξ_{SNR}	SNR criterion for parameterised data
$\hat{\xi}_{\text{SNR}}$	SNR criterion omitting $C^{\perp}_{\text{opt}}$
$\hat{\xi}^{\text{rez}}_{\text{SNR}}$	reciprocal SNR criterion omitting $C^{\perp}_{\text{opt}}$
$\hat{\xi}^{+}_{\text{SNR}}$	"sum-SNR" criterion omitting $C^{\perp}_{\text{opt}}$
$\hat{\xi}^{*}_{\text{SNR}}$	"product-SNR" criterion omitting $C^{\perp}_{\text{opt}}$
$\Phi(x)$	transformation function for SVM
ϕ_{VC}	VC-confidence
ρ	correlation coefficient
σ	singular value
σ^2	variance of a Gaussian distribution
$\sigma_d(x)$	decision function for SVM
$\theta(x)$	step function
$\Psi[n]$	mother wavelet
ν	degree of freedom
ω	continuous frequency variable

List of Figures

List of Tables