UNIVERSIDAD EAFIT

# The classification problem in machine learning: An overview

## With study cases in emotion recognition and music-speech differentiation

**MASTER IN ENGINEERING**

**17/07/2015**

This work addresses the well-known classification problem in machine learning. The goal of this study is to approach the reader to the methodological aspects of the feature extraction, feature selection and classifier performance through simple and understandable theoretical aspects and two study cases. Finally, a very good classification performance was obtained for the emotion recognition from speech.

# Contents

# List of Figures

# List of tables

# 1. Introduction

"The field of **Machine Learning** seeks to answer the question: How can we build computer systems that automatically improve with experience, and what are the fundamental laws that govern all learning processes?" (Mitchell, 2006)

As can be seen by the above definition Machine Learning is a broad field, it encompasses a variety of statistical techniques, computer science algorithms and heuristics. All of these try, in one way or another to answer that question.

Within the Machine Learning the specific area that will be this work's interest is the classification task. Here is where we have features or measurements as inputs to our model and produce a class or category that these sets belong to. A simple example would be to have a model that tells you when an apple is ripe or not (categories) based on its weight, color and density (measurements).

This field has a great number of applications in many areas (engineering, medicine, biology, psychology, economics, etc.) and they consist of several methods and techniques with publications in the area growing from around 152 publications in 1988 to 8494 in 2013 (source: Scopus). Clearly the area is current and rapidly growing.

The general outline in a task of this nature consists of three usual steps: The first and most crucial is feature generation and feature selection, where you preprocess the raw data to extract the features that your classifier will use. Its objective is to get features that clearly differentiate your classes and bring out the patterns in the data. The second step is selecting a classifier suitable for the task. Finally we have the training and validation steps where the method will learn from the data and then its performance will be evaluated. These steps can also be summarized in three questions: Which will my inputs be? What method should I use? How does the method perform?


**Figure 1 Steps for Classification task**

The questions above are not always easy to answer and lots of work can be put into each one, none the less they will be addressed throughout the following chapters.

While machine learning has encompasses lots of methods, not all of them perform well for all problems so it is important to establish which of them work well for a given case. The methods that will be explored in this work are: Linear Discriminant Analysis (LDA), Perceptron, Support Vector Machine (SVM), Naive Bayesian Classifier, Artificial Neural Networks (ANN), Classification Trees and kernel classifiers. While there will be a brief introduction to each of these methods the goal is not to dwell on each technique but rather to contextualize them with two audio databases.

The mentioned databases are the Berlin Emotional Speech database and a music speech database. In the first one the classifier has to differentiate between emotions in a speech signal, in the second one the idea is to classify a signal between music and speech.

## 1.1.Justification

While there are a lot of works and books in the general area of classification all of them take either an educational or an investigative approach. This work aims to fill the gap for a simple introduction to the subject, with the whole methodology being applied to the study cases, thus illustrating its use.

Both study cases represent interesting applications for the whole classification methodology and are current as well as challenging.

The case of music/speech differentiation represents the first step in an analysis of the information present in speech, which is crucial for human-machine interaction, and in music it can help in its automatic managing (Duan, Zhang, Roe, & Towsey, 2012).

Emotion recognition in speech presents a challenge both in its modelling as well as its characterization. Emotions convey non-linguistic information that alters the meaning of that is being said thus their automatic recognition allows for a more natural human-machine interaction, with applications ranging from avoiding traffic accidents to patient diagnoses (Koolagudi & Rao, 2012).

## 1.2.Objective

This work's intention is to introduce the field of classification in machine learning, going through all the steps in its methodology, starting with databases and feature extraction, then moving to feature selection and the classifiers themselves. All the methodology is introduced and presented with the accompanying study cases and at the end a more in-depth view of the emotion recognition case.

The goal of illustrating the whole methodology and presenting it with a study case is meant for the reader to take as an introduction to the subject of classification. Giving tools for the reader to come to its own conclusions as to what method to use in which case and planting the ground for both study cases to be further developed.

## 1.3.Outline

Each of the following chapters tackles a part of the classification task and gives a contextualization of it in the study cases for this work:

- **Chapter 2** Introduces the concept of features, their extraction and preprocessing, giving examples in the study cases yielding big groups of features for each of them.

- **Chapter 3** gives an overview of the feature selection and the performance measures for classifiers. As for the feature groups from the previews chapter selection is performed in several ways, yielding a total of seven final sets of features.

- **Chapter 4** provides an outline for two dimensionality reduction techniques using one of them with the study cases to obtain a two dimensional input vector.

- **Chapter 5** contains a short description for each of the classifiers and its results with all of the databases from the previews chapters, these results are given in a graphical way as well as a numerical one.

- **Chapter 6** is the final chapter where an application of the whole process for the entire Berlin database is presented; this is meant both as an example and a more specific case for an in-depth application of the whole classification methodology.

- **Chapter 7** provides the concluding remarks for this work.

# 2. Feature extraction

Features constitute the data which will be the input for the actual classifier; they can be as simple as raw data or as complex as the result from a previous process or system. It can be argued that they constitute the single most important part in the classification process; a good set of features can work well with a simple classifier whereas a bad one does not necessarily work with a complex classifier.

This section contains an overview of what feature extraction is. Starting with data bases and their use, following with the actual measurement and feature extraction from the data base, next the preprocessing of the data is discussed and finally the application of this concepts in the study cases.

## 2.1. Databases

A database consists of all the data that has been collected from whatever is going to be classified. It needs to follow an experimental design as much as possible; that is, having well defined factors (classes), dependent variables that are measured, control cases, distribution of the factors among the data, etc.

The two cases that will be studied in this work come from the Berlin Database of Emotional Speech (Burkhardt, Paeschke, Rolfes, Sendlmeier , & Weiss, 2005) and the GTZAN music/speech collection (Cook, 2015). The first consists of audios of actors who interpret 7 different emotions, the audios are sampled at 16kHz and last an average of 2.7s; the objective is to be able to classify these emotions. The second consists of audios containing either music or speech, these are sampled at 2050Hz and last around 30s; the objective is to be able to differentiate the two classes.

## 2.2. Features or measurements

The measurements are the features that are extracted either from the dataset of from the experiment itself. The features can be statistical data (Lambrou, Kudumakis, Speller, Sandler, & Linney, 1998) i.e. how often something happens in the dataset, how a certain measure is distributed or other indicators. Other alternative can be frequency analysis (Yang, Van Vuuren, & Sharma, 2000) (Ibrahim, Ambikairajah, Celler, & Lovell, 2007 ) or wavelet related features (Sun, Bebis, & Miller, 2002) (Karkanis, Iakovidis, Maroulis, Karras, & Tzivras, 2003) (Park, Choi, Nah, Jang, & Kim, 2008).

More specifically for audio signals there are four main kinds of features: low-level signal parameters (root-mean-square, bandwidth, zero-crossing rate, pitch, etc.), mel-frequency cepstral coefficients, psychoacoustic features (perception of temporal envelope modulations, sensation of signal strength, relative strength of high-frequency energy, etc.) and auditory filterbank temporal envelopes (a model representation of temporal envelope processing) (McKinney & Breebaart, 2003).

Any kind of feature that describes changes in the data can be susceptible to being used but it is highly dependent in the classification problem at hand, i.e. treating your data as a time series when it is clearly not one or extracting features from a model representation of a system that isn't the one you are working on.

## 2.3. Preprocessing

After the actual features have been extracted from the data at hand comes the preprocessing, this stage is meant to clean the data so it works better when it is processed. There are three main ways to preprocess the data: outlier removal and two kinds of normalization.

*Outlier removal* consists of identifying the data points which are too far away from the rest of the data these can cause errors during the training. There are three fundamental approaches, one is similar to unsupervised clustering, another models both normality and abnormality and is closer to classification, the last one models normality and is semi-supervised in nature (Hodge & Austin, 2004). In one way or another all of them assume an underlying structure to the data and some way to measure if a data point deviates in a great manner from the structure of the rest. An example would be to assume your data has a somewhat normal distribution and to remove all the points that are 3 or 4 standard deviations away from the mean.

The first kind of *normalization* consists of rescaling all the data so it stays either between [-1, 1] or [0, 1]. There are two main reasons to rescale a feature's values: One is so the classifiers don't get biased toward the features with the largest values, the other is so that the optimization methods such as gradient descent converge faster.

The second kind of *normalization* is standardization; in this case the each feature is scaled so that it has mean zero and unit variance.

## 2.4. Study case

Given the two databases presented in section 2.1 the extraction of features will focus in audio signals. For the case of emotion recognition from speech an overview is presented in (Ververidis & Kotropoulos, 2006) as well as (El Ayadi, Kamel, & Karray, 2011) and (Giannakopoulos & Pikrakis, 2014) covers audio processing in general.

This section will present an overview of the features extracted from the databases, first giving an outline of wavelets and then presenting the actual features.

### 2.4.1. Multiresolution analysis

This analysis refers to the splitting of the original signal into a hierarchy of signals of increasing detail; this is done with the discrete wavelet transform (DWT) (Mallat, 1989). Each hierarchical level contains different information from the original signal and finer or coarser details of the signal.

The resulting levels of decomposition can be treated as new signals that result from the original one and extract features from them.

### 2.4.2. Extracted features

The Berlin database is resampled to 8 kHz to better resemble telephone quality speech, which would be more readily available in general, and divided in windows, while (Lu, Zhang, & Jiang, 2002) use windows of 1 second, others use other sizes (Tzanetakis, Essl, & Cook, 2001) (Scheirer & Slaney, 1997), in this case the windows have a length of 1 second and an overlap between windows of 0.4 seconds. These make up the total number of audio samples that will be used to extract the final features. It is worth noting that for the study case only two emotions will be classified, those are sadness and anger, the main reason for this choice is their easier differentiation.

The music/speech database is resampled to 8 kHz, so that it's conditions are closer the other study case, and used as is; each audio of 30 seconds will constitute one sample from which the features will be extracted.

From both sets of audios a multiresolution decomposition is performed to up to 10 levels, this approach has been previously used (El Ayadi, Kamel, & Karray, 2011) and gives a wider ranges of features from which to perform the selection stage later on. The procedure is done with the Daubechies 1, 6, 8 and 10 wavelets. Performing the decomposition leaves each sample with a new set of 11 decomposed signals plus the original one. After the decomposition from each signal the 24 features presented in Table 1 Extracted features are measured. This makes a total of 1072 features after extracting the ones that don't yield any results in a certain level.

| Features | |
|---|---|
| **1** | Mean |
| **2** | Variance |
| **3** | Energy |
| **4** | Shannon entropy |
| **5** | Maximum value of the PSD |
| **6** | Frequency with the maximum value in the PSD |
| **7** | Fourier entropy |
| **8** | First coefficient of an order 2 AR (AR1) |
| **9** | Second coefficient of an order 2 AR (AR2) |
| **10** | RMS |
| **11** | IAV |
| **12** | Wavelength |
| **13** | Zero crossing rate |
| **14** | Pitch |
| **15** | Minimum of the absolute value of the spectrum |
| **16** | Maximum value of the absolute value of the spectrum |
| **17** | Standard deviation of the absolute value of the spectrum |

| | |
|---|---|
| **18** | Skewness |
| **19** | Kurtosis |
| **20** | Inter quartile range (IQR) |
| **21** | Inter quartile range of the absolute value of the spectrum |
| **22** | Kurtosis of the absolute value of the spectrum |
| **23** | Mean of the absolute value of the spectrum |
| **24** | Skewness of the absolute value of the spectrum |

**Table 1 Extracted features which are commonly used in signals**

# 3. Feature selection

After the extraction it is necessary to evaluate which of them actually differentiate the classes. This section explores the process of feature selection which consists of identifying a subset of features that best differentiates the classes so that the classifier will perform better. The two main purposes are to get the smaller set of features and the best discrimination between the classes, ideally to have separable classes so that the classifier can aspire to 100% accuracy.

## 3.1. Class differentiation measurements

There several ways to measure how well a certain feature can discriminate between classes, some of them are the divergence, Chernoff Bound and Bhattacharyya Distance, the receiver operating characteristic curve (ROC) and scatter matrices (Theodoridis & Koutroumbas, 2009). This section covers the last two, which are the ones employed in the study case for feature selection.

### 3.1.1. ROC

The receiver operating characteristic curve measures the overlap between the pdfs of two classes for a respective feature. The ROC measures the area under the curve (AUC) for those overlaps. When the feature discriminates perfectly the separation has a value of 0.5 and when the overlap is thorough then the value is 0 (Fawcett, 2006).

### 3.1.2. Scatter matrices

Scatter matrices intend to approximate the mean of each class from their respective global value (Theodoridis & Koutroumbas, 2009), the main idea is to be able to measure when the classes are clustered around their means and separated among themselves.

The within-class scatter matrix is measured, which incorporates the probability of a certain class in the dataset as well as the covariance of that class. The between-class scatter matrix serves as an approximation of the mean of each class to the global mean value. The sum of these two matrices is known as the mixture scatter matrix which corresponds to the covariance matrix of the feature vector with respect to the global mean.

It is worth noting that the scatter matrices criteria take a special form in the one-dimensional case in a two class problem; here the Fisher's discriminant ratio is obtained. This ratio has a multiclass case were averaging forms are used (Theodoridis & Koutroumbas, 2009).

An advantage of using scatter matrices as the discrimination criteria is that it is simpler and easily computed. This kind of measure works both with multiclass cases and feature vectors, which constitutes another benefit when compared to other methods.

## 3.2. Classification performance measurement

This section covers some of the usual performance measures for a given classifier and their interpretation. A more applied view will be shown in the classifiers section where these measures will give an idea of how each of the shown classifiers performs with each dataset.

### 3.2.1. Confusion matrix

A confusion matrix is just a contingency table that helps visualize the performance of a given algorithm. The columns represent the predicted classes and the rows the actual classes of the data. The confusion matrix can also be used to illustrate the following measures.

### 3.2.2. Accuracy

The accuracy of a classifier is defined as the percentage of data that was classified correctly.

### 3.2.3. Sensitivity and specificity

The sensitivity (also called recall) and specificity are used in a two-class problem. The sensitivity corresponds to the correctly classified members of the first class with respect to the total number of members of this class. The specificity is the same but for the members of the second class. They are also called the true positive rate and the true negative rate.

### 3.2.4. Precision

The precision measures the rate of labeled members of a class actually belong to that class.

### 3.2.5. F1 score

The F1 score also called F score or F measure is a balance between precision and recall. It is defined as $2(precision * recall)/(precision + recall)$ and was introduced in (van Rijsbergen, 1979).

## 3.3. Scalar selection

The scalar selection of features is the general case where features are tested individually for the differentiation power they have for a given dataset. A way to do this is by taking a class differentiation measure and ranking the features from the most discerning to the least. After this ranking a composite ranking can be performed where weights are given to both the differentiating power of a feature, as well as the correlation it has with the top ranked features. The later composite ranking intents to ensure that the top ranked features don't present a high correlation.

## 3.4. Feature vector selection

The selection of a feature vector is performed as to find the best subset of features. The methods under this category are search heuristics that evaluate how well a certain subset performs; this constitutes the main difference with the scalar selection where the features were tested individually. One way to perform a feature vector selection is to do an exhaustive search of all the subsets of features, clearly this takes great amounts of time.

There are other methods that perform a suboptimal search such as sequential forward and backward selection as well as forward floating search selection (Pudil, Ferri, Novovicova, & Kittler, 1994) (Zongker & Jain, 1996). Other more general heuristics can also be used if well-defined such as genetic algorithms, randomized searches, etc.

## 3.5. Study case

Feature selection was performed over the two datasets of 1072 features each one from the databases mentioned in the Feature extraction section (2.4). First preprocessing was performed with the purpose of cleaning up the data, taking away features that had the same values as well as features that didn't present any changes at all. These are the features that have the same value for all cases that have the same values as another feature or that couldn't be measured.

### 3.5.1. Music/Speech dataset

From the music/speech dataset after preprocessing there were 960 features. From those a sequential forward floating selection was performed using scatter matrices as the differentiation measure, this was done once looking for a subset of 3 features that differentiate the two classes in a satisfactory manner (M/S). The resulting features correspond to the ones in Table 2 Final features for the M/S database.

| Music-Speech (M/S) database features | |
|---|---|
| **Multiresolution analysis information** | **Feature** |
| Db6 level 10 of detail | 13 |
| Db8 level 5 of detail | 21 |
| Db8 level 10 of detail | 7 |

**Table 2 Final features for the M/S database**

### 3.5.2. Berlin dataset

The Berlin dataset ended up with 970 features after the preprocessing. From that set of features several selection methods were performed. The first method corresponds to a sequential forward floating selection of 2 features using scatter matrices as the differentiation measure (Br.1), this dataset is shown in Figure 3 Sequential forward floating selection of two features (Br.1). The same method was also used to obtain a subset of 4 features (Br.2). The general process for the treatment of the data is presented in the following chart (Figure 2 Data treatment process).

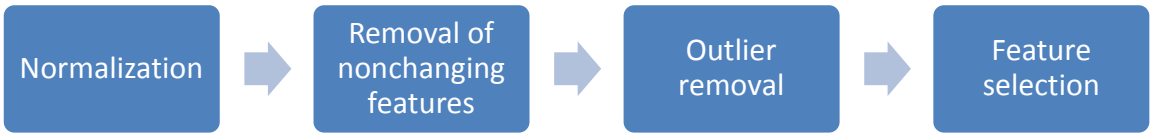



**Figure 2 Data treatment process**

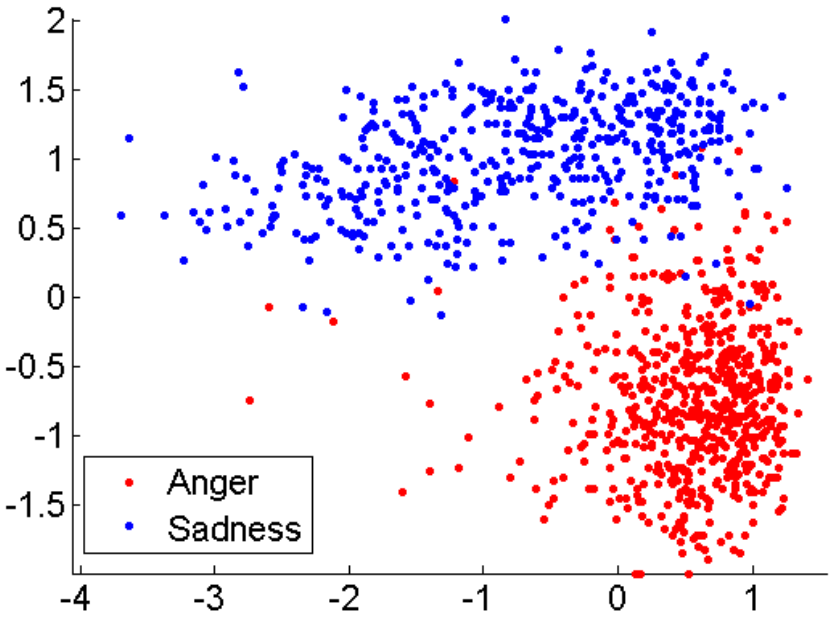



**Figure 3 Sequential forward floating selection of two features (Br.1)**

A composite feature ranking using ROC as differentiation measure is performed giving a weight of 0.3 to the class discerning measure and a 0.7 weight to the correlation measure. From this ranking the top two features were selected (Br.3) which is shown in Figure 4.

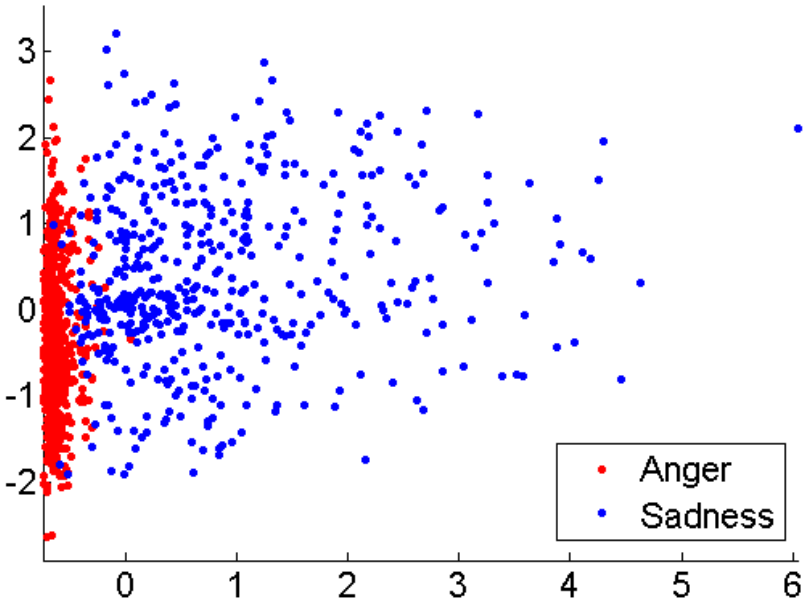



**Figure 4 First two features of the composite feature ranking (Br.3)**

Finally with no selection performed the energy of the signal and the zero crossing rate conform the final selected dataset (Br.4) which is shown in Figure 5. The features of all of the berlin datasets appear in Table 3 Features for each of the four Berlin database's selectionsTable 3.

| Dataset | Multiresolution analysis information | Feature |
|---|---|---|
| Br.1 | Db6 level 4 of detail | 11 |
| | Db6 level 4 of detail | 12 |
| Br.2 | - | 18 |
| | Haar level 5 of detail | 18 |
| | Db6 level 4 of detail | 12 |
| | Db6 level 8 of detail | 11 |
| Br.3 | Haar level 6 of detail | 13 |
| | Db6 level 9 of detail | 1 |
| Br.4 | - | 3 |
| | - | 13 |

**Table 3 Features for each of the four Berlin database's selections**

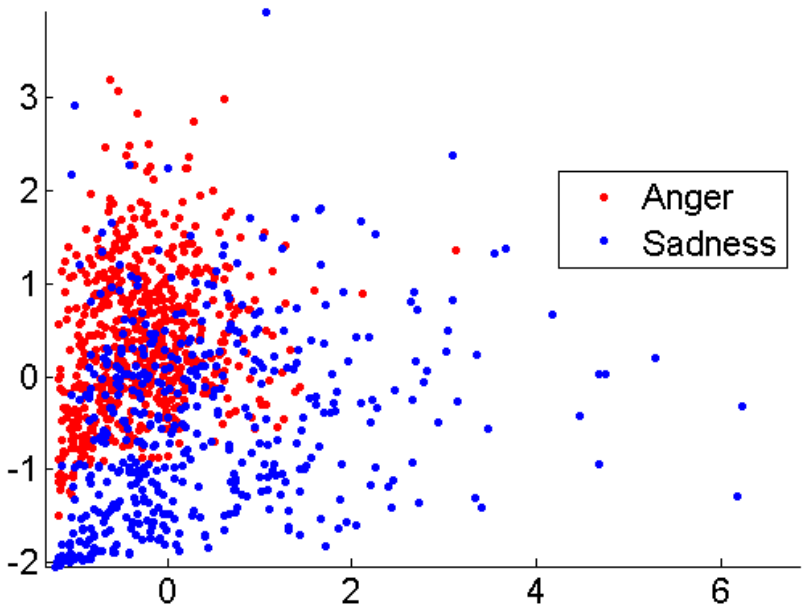



**Figure 5 Energy of the signal and the zero crossing rate (Br.4)**

# 4. Dimensionality reduction

The process of dimensionality reduction, as the name suggests, aims to express the information contained in the features in less dimensions than the original set. These techniques use linear and nonlinear transformations of the data to achieve the goal in some optimal way.

## 4.1. Principal component analysis

Also known as PCA (Jolliffe, 2002), principal component analysis starts from the original set of features and applies a linear transformation so that the components of the resulting set are uncorrelated. Afterwards the most significant of those components are chosen.

## 4.2. Fisher's linear discriminant analysis

Fisher's LDA (Welling, 2005) is similar to PCA, the main difference being that this is a supervised model, meaning that the class the data belong to is taken into account. The goal is for the means of the data to be as far apart as possible and for the variance to be as small as possible.

## 4.3. Study case

For the previously mentioned datasets there some dimensionality reduction was used to obtain better visualization of the separation between the classes.

### 4.3.1. Music/Speech database

The M/S features live in a three dimensional space, from these three final datasets were obtained. The first and second are projections of the data into a 2D plane, the projection is not an optimal one it is done in an empirical manner the sets are labeled M/S.P2 and M/S.P3, Figure 7 and Figure 8 respectively. The third is the result of the projection done by a PCA in a 2D space labeled M/S.P1 and is shown in Figure 6.

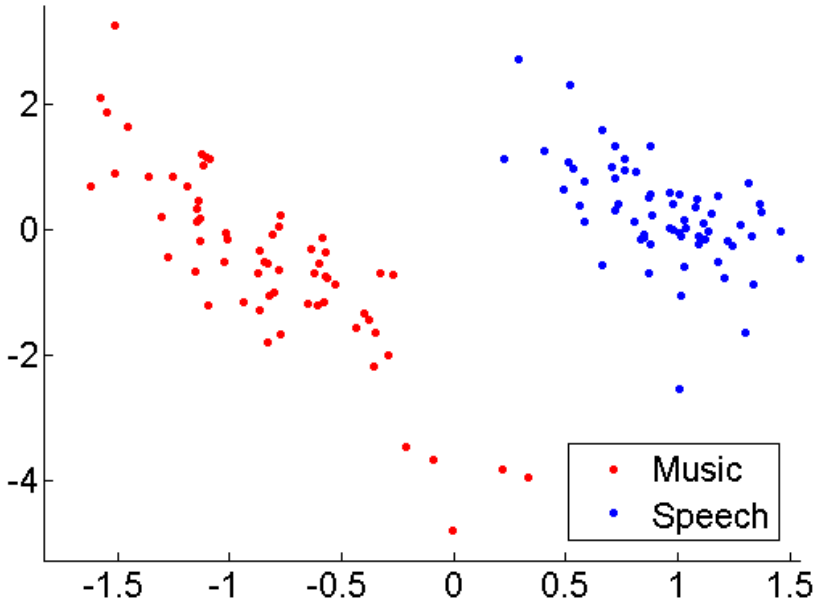



**Figure 6 PCA projection of the M/S database selected features with linearly separable classes (M/S.P1)**

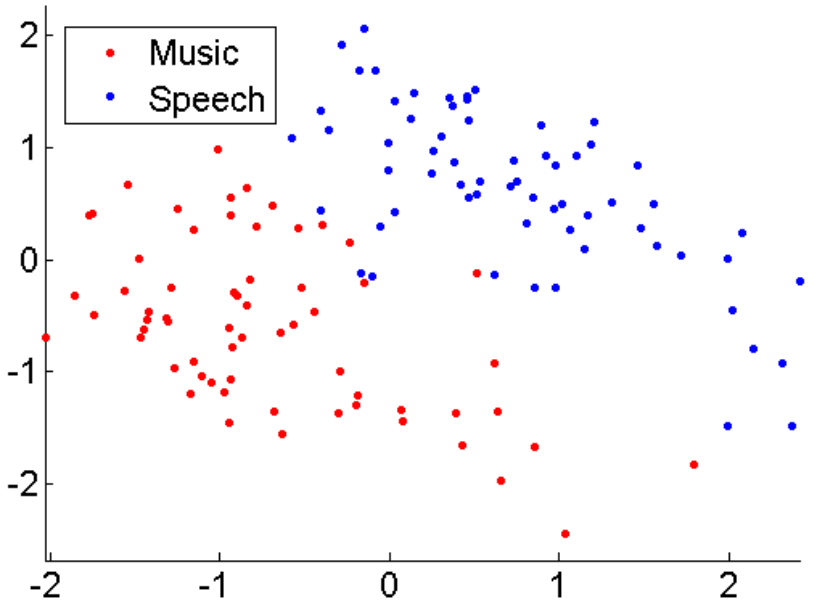



**Figure 7 2D projection of the M/S database selected features (M/S.P2)**

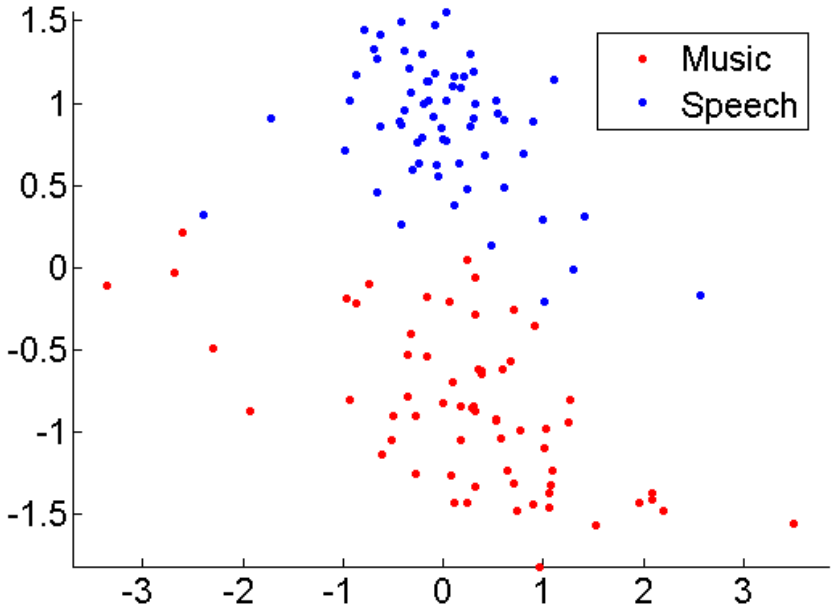



**Figure 8 2D projection of the M/S database selected features with non-linearly separable classes (M/S.P3)**

### 4.3.2. Berlin database

From the Br.2 features a PCA was performed so that a 2D projection is obtained. The final Br.2 set is the one that will be used and is shown in Figure 9.

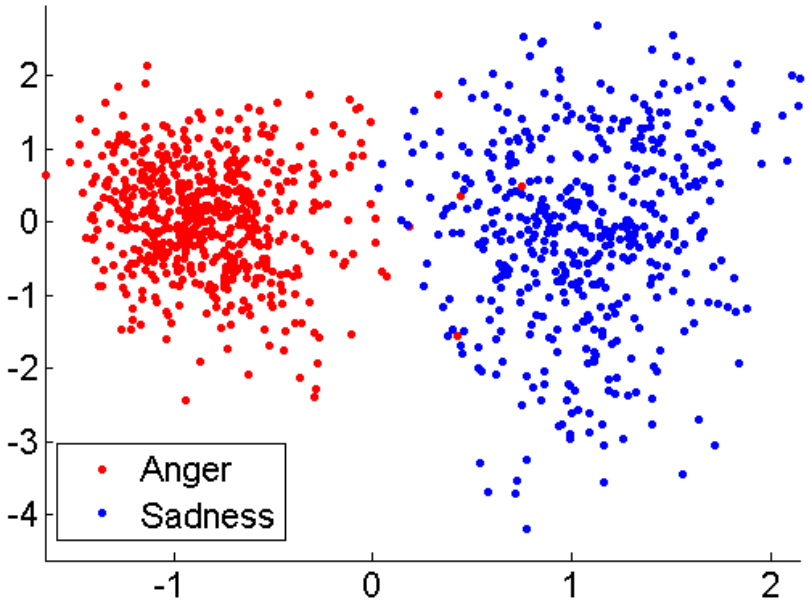



**Figure 9 PCA projection of a four feature sequential forward floating selection (Br.2)**

# 5. Classifiers

After the final set of features is obtained the training of the classifier can be performed. This section covers nine classifiers which are split into three groups: linear, nonlinear and kernels. Along with an introduction to each classifier its overall performance in the seven datasets obtained in previous sections is shown. These datasets represent a mixture of conditions from the perfect scenario where the classes a linearly separable to one where they are almost completely mixed. The sets provide a good view to the actual performance of each classifier.

For a more in-depth overview of the classifiers introduced in this section go to (Theodoridis & Koutroumbas, 2009) or to (Friedman, Hastie, & Tibshirani, 2001).

## 5.1. Data partitioning and training

When using data, and after it has been processed it must be divided into at least two: training and testing subsets. An optional set would be the validation subset. There are many ways to make this division, which on to use depends a lot in the kind of problem that is being dealt with. The data in each set is selected either at random or in a way that is consistent with the nature of the problem.

This partitioning and its size has been explored in several woks such as (Foody, Mathur, Sanchez-Hernandez, & Boyd, 2006), (Foody, McCulloch, & Yates, 1995), (Kohavi, 1995), (Arlot & Celisse, 2010).

The training set consists of the data that the classifier is going to be trained with, in other words, the data that it is going to know. This set is usually the biggest one consisting of around 60 and 80 percent of the total data. The size of the set can vary depending on the classifier that is being used (some are more sensitive to the amount of data), on the amount of data available or other factors.

The validation set works as a secondary training set that is mostly used to tune parameters on the classifier or to stop training so that the classifier is not over fitted to the training set.

Finally the testing set is the one where the classifier is tested to know it's accuracy in a closer scenario to the real case. The data in this set shouldn't be used in any kind of training or tuning of the classifier.

## 5.2. Linear classifiers

The classifiers in this category use a hyperplane as the boundary of classification, that is to say that the boundary they define is a line in 2D and a plane in 3D, this will become clear in the applications of each of them. It is important to note that a linear classifier can only have 100% accuracy when the classes are linearly separable.

### 5.2.1. LDA

A Linear Discriminant Analysis (LDA) classifier is a method that models class density as a Gaussian distribution. We have the class $k$ density given by a multivariate Gaussian

$$f_k(x) = \frac{1}{(2\pi)^{\frac{p}{2}}|\Sigma_k|^{\frac{1}{2}}} e^{-\frac{1}{2}(x-\mu_k)^T\Sigma_k^{-1}(x-\mu_k)} \qquad (1)$$

Where $\mu_k$ is the class mean and $\Sigma_k$ the covariance matrix.

The LDA is the special case when the covariance matrices are assumed to be the same for all classes $\Sigma_k = \Sigma \; \forall$k, this results in the classifier being a linear classifier which can be seen by looking at the log-ratio of the two classes

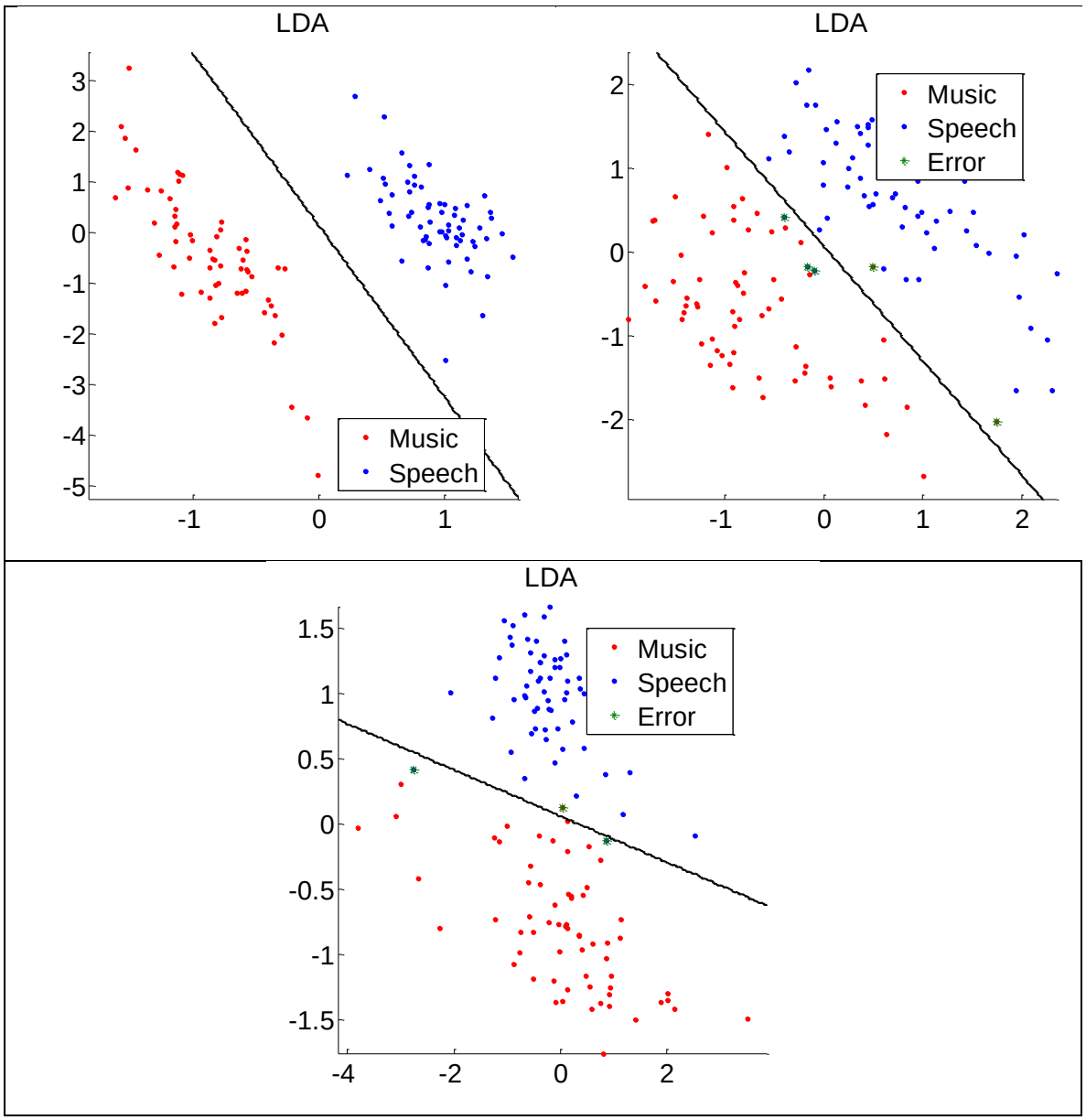
$$\begin{aligned} \log\frac{\Pr(G=k|X=x)}{\Pr(G=l|X=x)} &= \log\frac{f_k(x)}{f_l(x)} + \log\frac{\mu_k}{\mu_l} \\ &= \log\frac{\mu_k}{\mu_l} - \frac{1}{2}(\mu_k+\mu_l)^T\Sigma^{-1}(\mu_k-\mu_l) \\ &+ x^T\Sigma^{-1}(\mu_k-\mu_l) \end{aligned} \qquad (2)$$

It is clear that this equation in linear in $x$, this implies that the decision boundary between the classes is linear (Duda, Hart, & Stork, 2012).

Figure 10 shows the results of classifying the databases with an LDA, the accuracy is shown in Table 4. The most interesting result is the classification of the Br.2 set, where the variances of the two classes are very different and so the resulting boundary from the LDA method is not the best one.

| Database | Accuracy |
|---|---|
| M/S.P1 | 100% |
| M/S.P2 | 96.1% |
| M/S.P3 | 97.7% |
| Br.1 | 98% |
| Br.2 | 87.1% |
| Br.3 | 77.7% |
| Br.4 | 99.3% |

**Table 4 LDA accuracy for each dataset**


LDA
Music
Speech
LDA
Music
Speech
Error
LDA
Music
Speech
Error

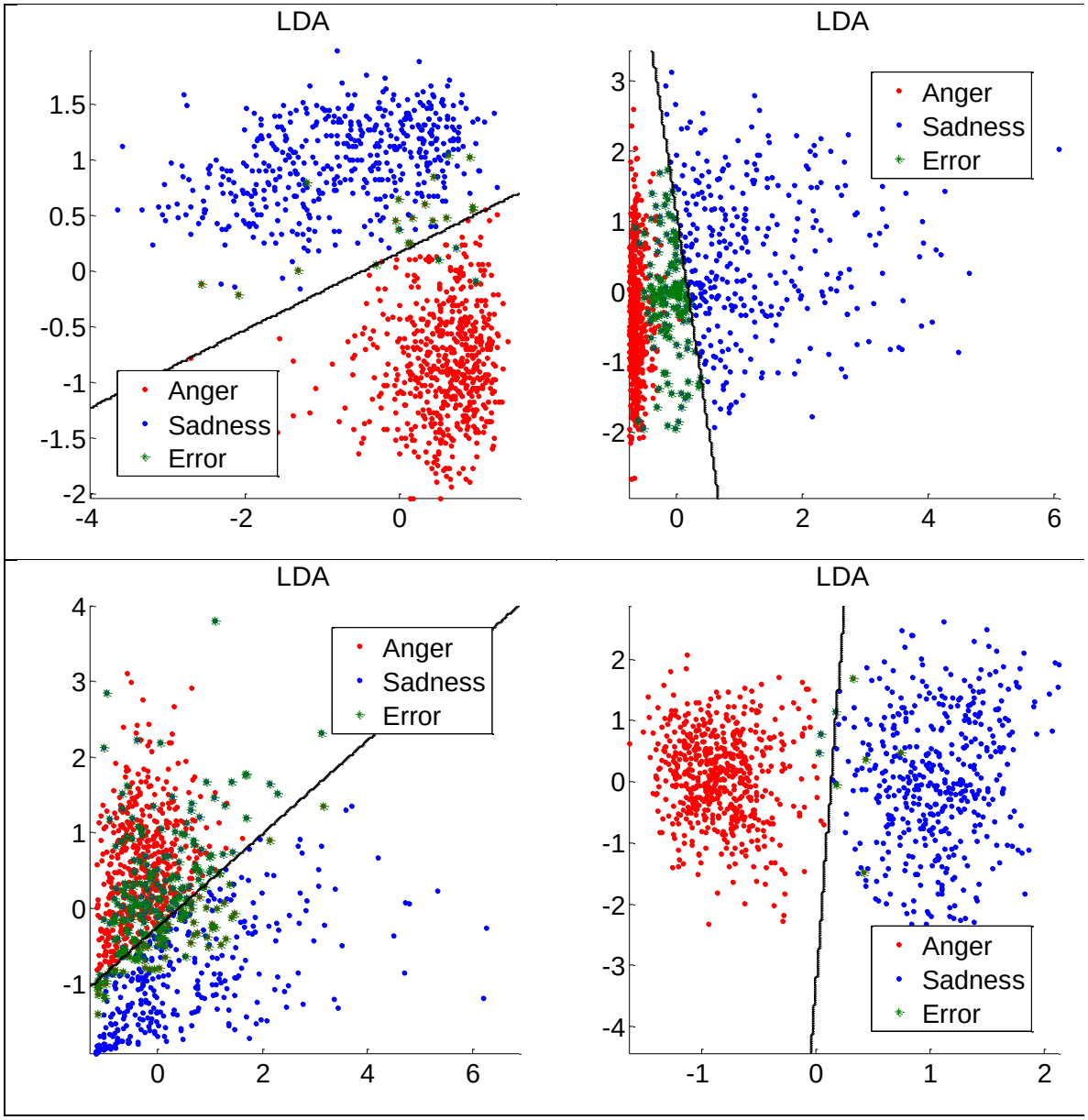



**Figure 10 LDA decision boundaries and classification results for M/S.P1, M/S.P2, M/S.P3, Br.1, Br.2, Br.3 and Br.4.**

### 5.2.2. Perceptron

To talk about the perceptron algorithm first the decision hypersurface in $l$-dimensional feature space must be introduced

$$g(x) = w^T x + w_0 = 0 \qquad (3)$$

Where $w = [w_1, w_2, \dots, w_l]^T$ is a weight vector and $w_0$ is known as the threshold. The perceptron algorithm assumes that there exists a hyperplane such that

$$w^T x + w_0 < 0 \quad \forall x \in k_2 \qquad (4)$$
$$w^T x + w_0 < 0 \quad \forall x \in k_2 \qquad (5)$$

Where $k_1$ and $k_2$ are two classes, now the parameters $w' = [w^T, w_0]$ need to be found. In this case the perceptron cost function is defined as

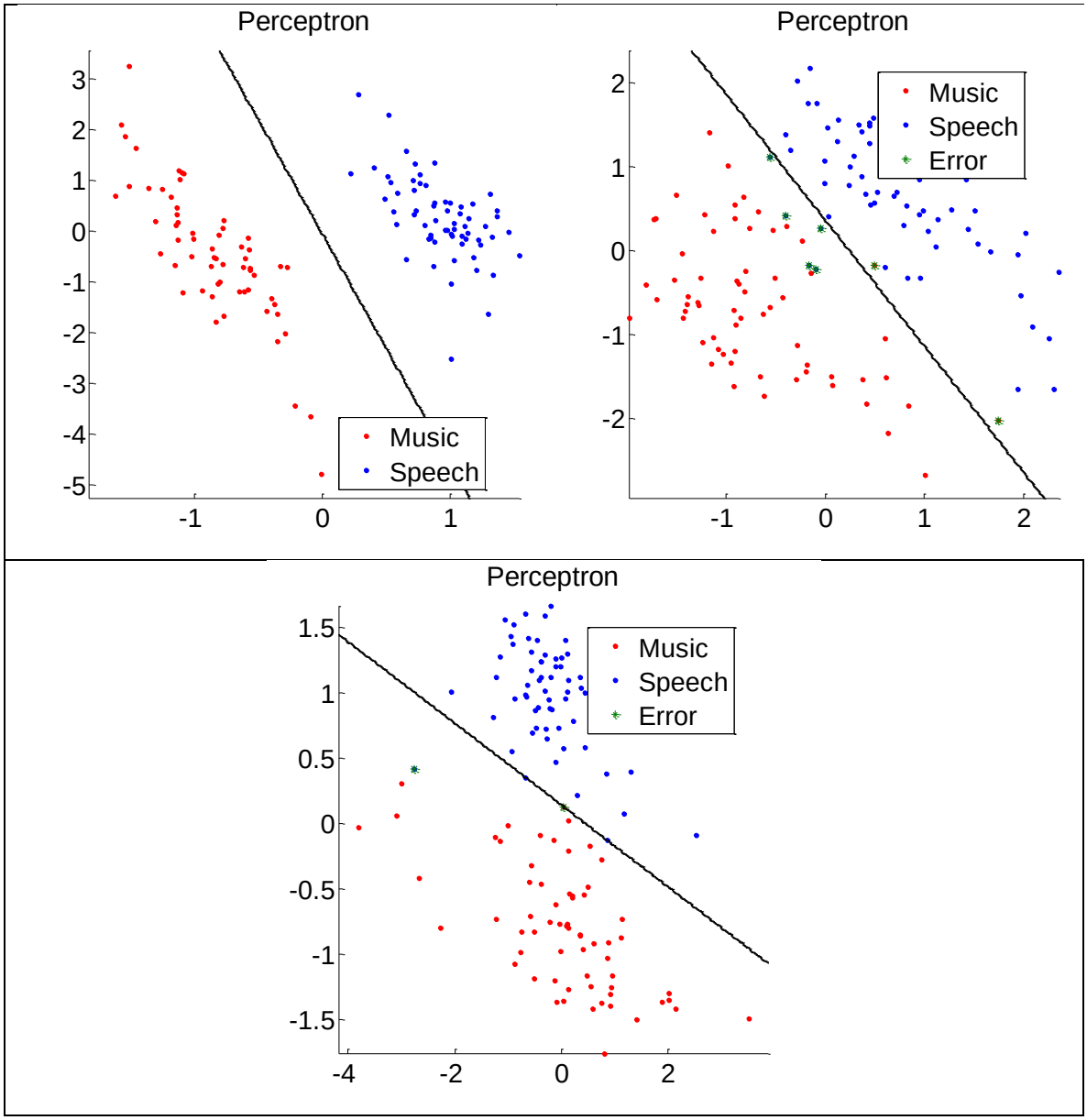
$$J(w) = \sum_{x \in Y} (\delta_x w^T x) \qquad (6)$$

Where $Y$ is the set of training vectors that the hyperplane is currently misclassifying given a specific set of parameters $w$. $\delta_x$ is defined as -1 if $x \in k_1$ and +1 if $x \in k_2$, this ensures that the sum is always positive and becomes zero when $Y$ is empty. This cost function is minimized iteratively with the gradient descent algorithm although many other methods can be used. (Rosenblatt, The perceptron: a probabilistic model for information storage and organization in the brain, 1958) (Rosenblatt, Principles of neurodynamics. perceptrons and the theory of brain mechanisms, 1961).

In Figure 11 the results of the perceptron classifier for all seven databases are presented and Table 5 shows the accuracies for them.

| Database | Accuracy |
|---|---|
| M/S.P1 | 100% |
| M/S.P2 | 94.5% |
| M/S.P3 | 98.4% |
| Br.1 | 98% |
| Br.2 | 96.4% |
| Br.3 | 78.3% |
| Br.4 | 99.3% |

**Table 5 Perceptron accuracy with each dataset**


Perceptron
Music
Speech
Perceptron
Music
Speech
Error
Perceptron
Music
Speech
Error

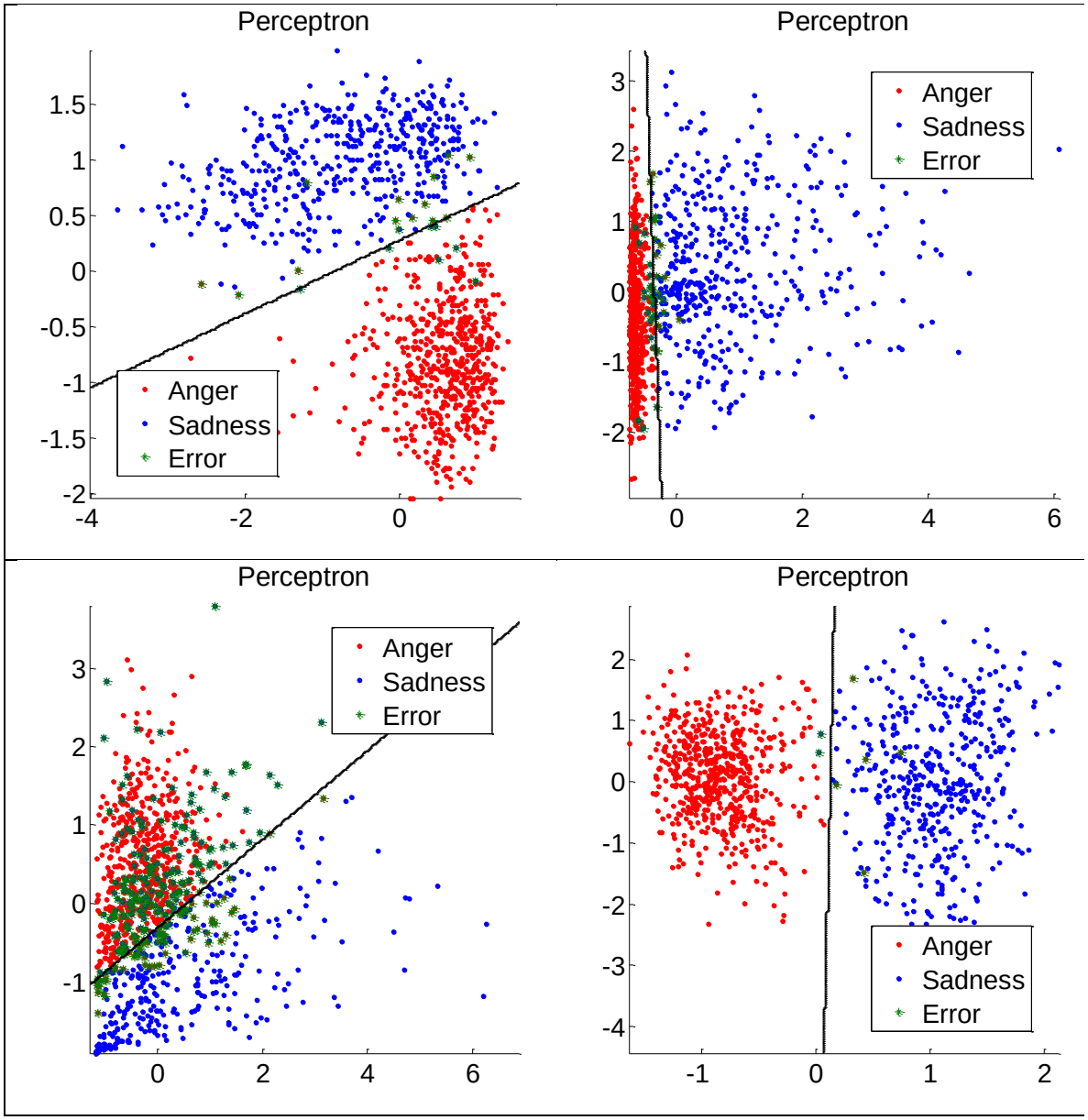



**Figure 11 Perceptron decision boundaries and classification results for M/S.P1, M/S.P2, M/S.P3, Br.1, Br.2, Br.3 and Br.4.**

### 5.2.3. Support Vector Machine

Another linear classifier is the support vector machine (SVM) which in the optimal case not only differentiates the classes but also keeps the decision boundary as far as possible from both so that the resulting classifier is more robust.

In this case we have two classes $k_1$ and $k_2$ with N feature vectors associated with each of them. The idea is to obtain a hyperplane ($g(x) = w^T x + w_0 = 0$) that classifies correctly the training set, but

this hyperplane is not unique and so SVM tries to find the one that leaves the biggest margin between the two classes.

For this purpose we know that a hyperplane is characterized by its direction $w$ and position $w_0$. The normalized margin between the classes and the hyperplane is restricted such that if it is 1 for one class, then it is -1 for the other, that is:

$$w^T x + w_0 \geq 1, \qquad \forall x \in k_1 \qquad (6)$$
$$w^T x + w_0 \leq -1, \qquad \forall x \in k_2 \qquad (7)$$

Denoting $y_i$ (+1 for $k_1$, -1 for $k_2$) the class that $x_i$ belongs to, then the SVM optimization problem will be:

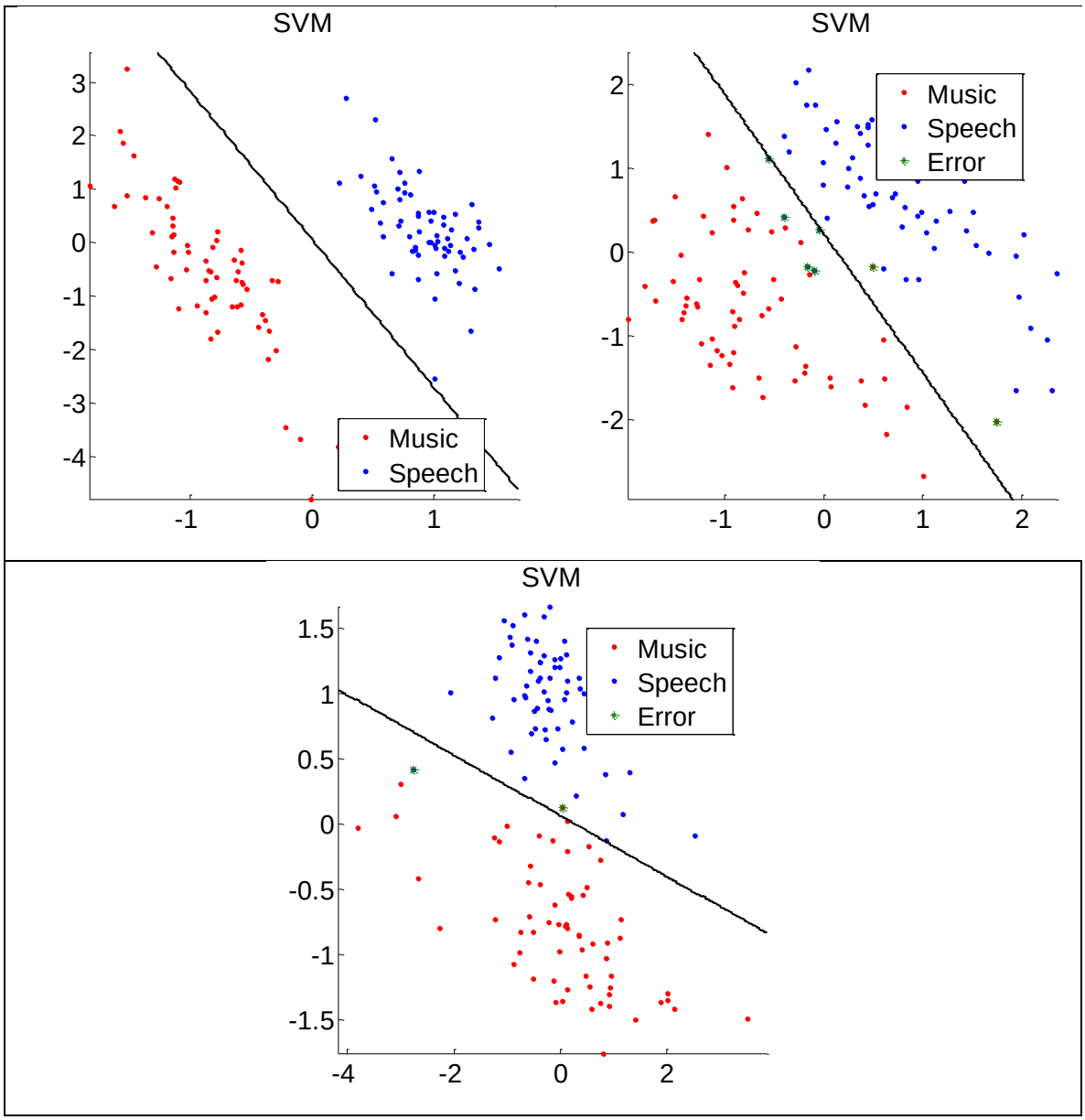
$$\text{Minimize} \quad J(w, w_0) \equiv \frac{1}{2}\|w\|^2 \qquad (8)$$
$$\text{Given that} \quad \mathrm{y_i}(w^T x_i + w_0) \geq 1, \quad i = 1,2,\dots,N \qquad (9)$$

Solving this problem will yield a unique hyperplane that is the one defining SVM. It is worth noting that for the case where the classes are not linearly separable this formulation changes and a parameter $C$ emerges which controls the importance of the error versus the margin (Vapnik, 2013).

Table 6 SVM accuracy in each of the datasets and Figure 12 show the results of an SVM classifier with parameter $C = 1$ for all seven databases.

| Database | Accuracy |
|---|---|
| M/S.P1 | 100% |
| M/S.P2 | 94.5% |
| M/S.P3 | 98.4% |
| Br.1 | 98.1% |
| Br.2 | 93.1% |
| Br.3 | 79.5% |
| Br.4 | 99.2% |

**Table 6 SVM accuracy in each of the datasets**


SVM
Music
Speech
SVM
Music
Speech
Error
SVM
Music
Speech
Error

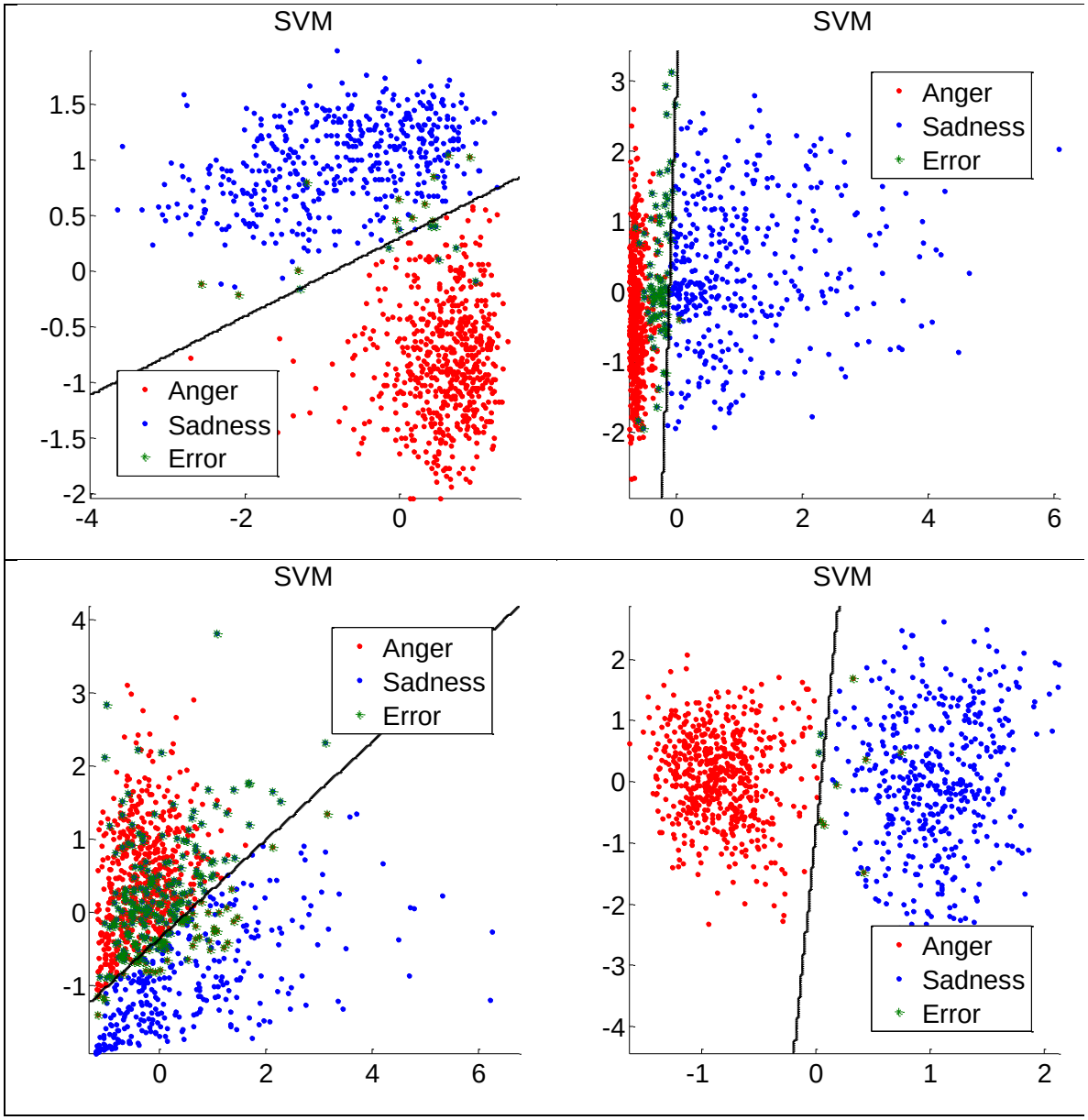



**Figure 12 SVM decision boundaries and classification results for M/S.P1, M/S.P2, M/S.P3, Br.1, Br.2, Br.3 and Br.4.**

A better approximation for the parameter $C$ for each database is presented in Table 7. Here either the accuracy or the margin between the classes is improved.

| Database | Accuracy | $C$ |
|---|---|---|
| M/S.P1 | 100% | 0.5 |
| M/S.P2 | 96.1% | 5 |
| M/S.P3 | 98.4% | 3 |
| Br.1 | 98.3% | 0.3 |
| Br.2 | 96.6% | 10 |
| Br.3 | 79.5% | 10 |

| | | |
|---|---|---|
| Br.4 | 99.2% | 1 |

**Table 7 Better approximation for SVM for each dataset and the corresponding parameter**

## 5.3. Non-Linear classifiers

The classifiers in this category use a hypersurface as the boundary of classification instead of a hyperplane.

### 5.3.1. Naive Bayes

This classifier's tries to classify an input in its most probable. In this case we assume that each class has a Gaussian distribution defined as:

$$p(x) = \frac{1}{(2\pi)^{\ell/2}|\Sigma|^{1/2}} e^{\left(-\frac{1}{2}(x-\mu)^T\Sigma^{-1}(x-\mu)\right)} \qquad (10)$$

Where $x$ is a vector of $\ell$ features, $\mu$ a vector of $\ell$ expected values, $\Sigma$ is the covariance matrix $\ell$x$\ell$ and $|\Sigma|$ its determinant. Given the exponential nature of the equation the following discriminant function is used:

$$g_i(x) = -\frac{1}{2}(x-\mu_i)^T{\Sigma_i}^{-1}(x-\mu_i) + ln\big(P(k_i)\big) + c_i \qquad (11)$$

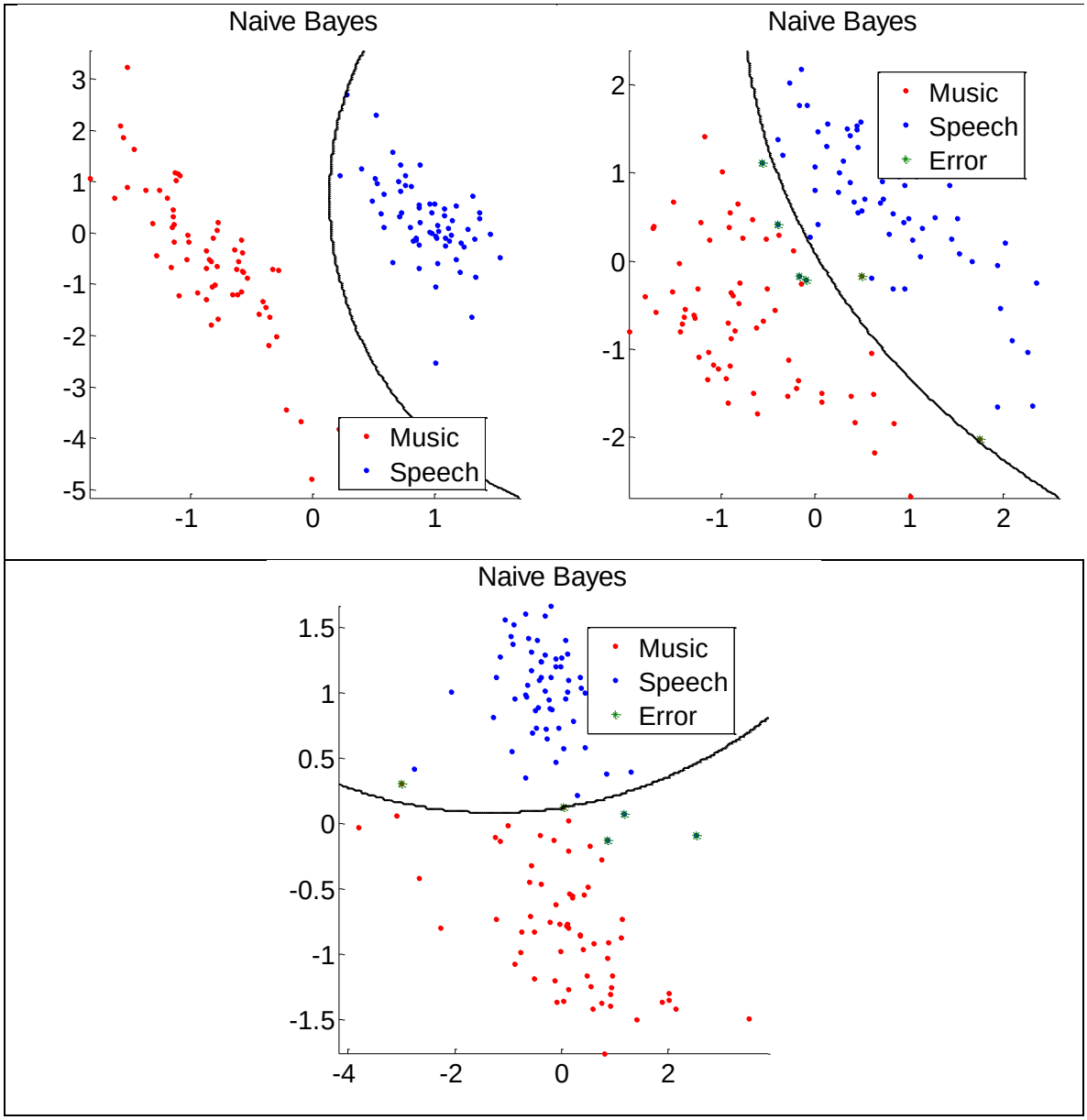
Where $P(k_i)$ is how probable class $k_i$ is.

For the Naive Bayes classifier it is assumed that the features are independent and so the parameters for each Gaussian distribution are easier to find.

The results for all seven databases are shown in Table 8 and Figure 13, the most notable here is the result in Br.2 where the classifier a boundary that captures well the relationship between the classes (Rish, 2001).

| Database | Accuracy |
|---|---|
| M/S.P1 | 100% |
| M/S.P2 | 95.3% |
| M/S.P3 | 96.1% |
| Br.1 | 97.8% |
| Br.2 | 96.5% |
| Br.3 | 78.8% |
| Br.4 | 99.3% |

**Table 8 Naive Bayes accuracy for each dataset**


Naive Bayes
Music
Speech
Naive Bayes
Music
Speech
Error
Naive Bayes
Music
Speech
Error

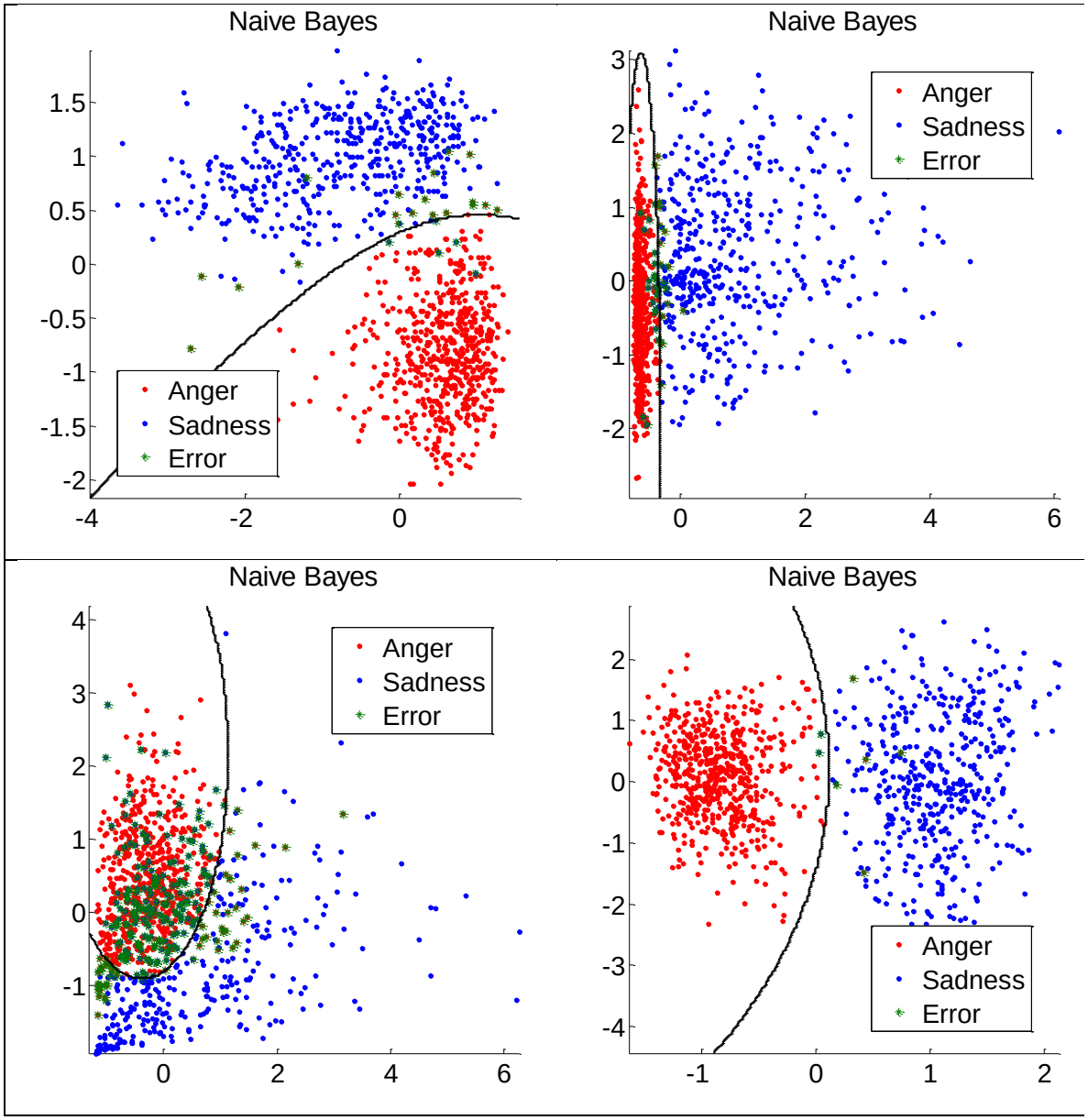



**Figure 13 Naive Bayes decision boundaries and classification results for M/S.P1, M/S.P2, M/S.P3, Br.1, Br.2, Br.3 and Br.4.**

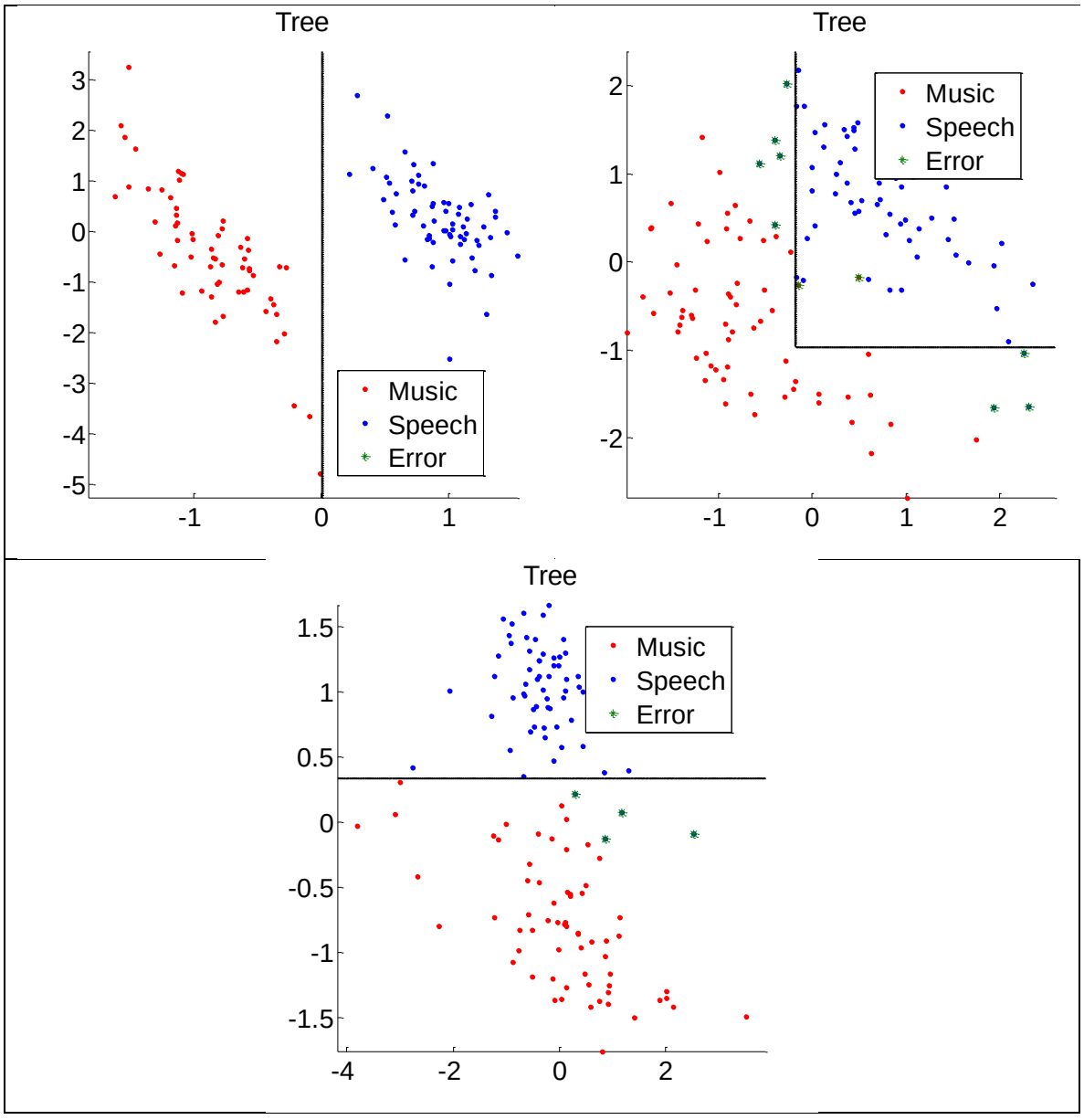
### 5.3.2. Decision tree

A decision tree can be viewed a step by step process where in every step the data is divided into a set of classes. In this tree-like hierarchical structure each rule represents a rule and there are several methods that deal with how to construct this rules and the tree. More information can be found in (Khan & Alpaydin, 2004).

The results in the seven databases of the study case are shown in Table 9 and Figure 14.

| Database | Accuracy |
|---|---|
| M/S.P1 | 98.4% |
| M/S.P2 | 92.2% |
| M/S.P3 | 96.9% |
| Br.1 | 98.3% |
| Br.2 | 98% |
| Br.3 | 86.4% |
| Br.4 | 99.6% |

**Table 9 Decision tree accuracy for each of the datasets**



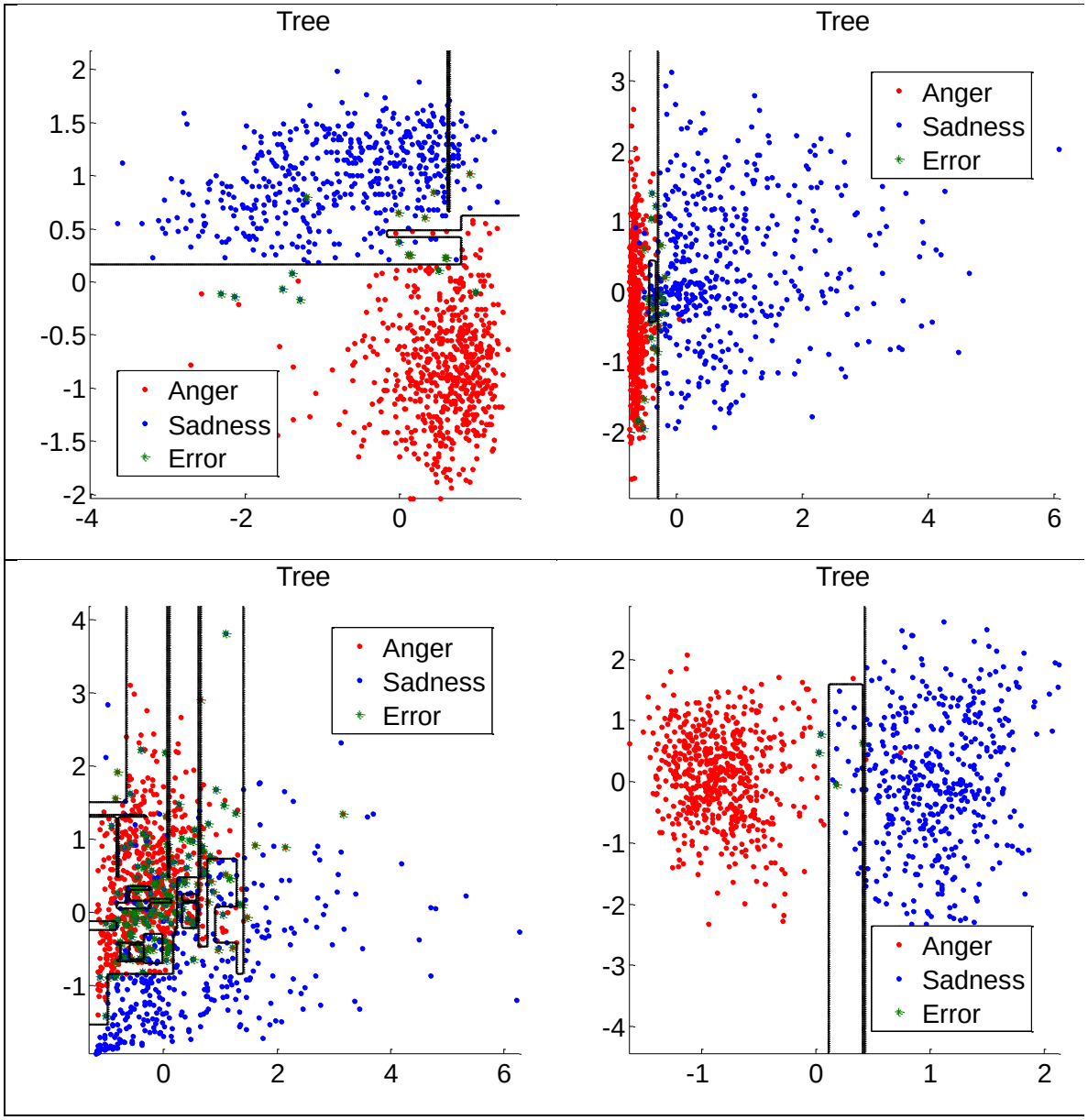



**Figure 14 Decision tree decision boundaries and classification results for M/S.P1, M/S.P2, M/S.P3, Br.1, Br.2, Br.3 and Br.4.**

### 5.3.3. Artificial Neural Network

An Artificial Neural Network (ANN) (Rosenblatt, Principles of neurodynamics. perceptrons and the theory of brain mechanisms, 1961) (Widrow, 1960) is in its basic form consists of several perceptrons (or Neurons) connected in a network topology, the neurons are arranged by layers with connections allowed between the nodes in successive layers. The connections between nodes are represented by a weight parameter. This architecture is highly nonlinear and is a black box model, meaning that after the training of the weights the result is not interpretable but usable.

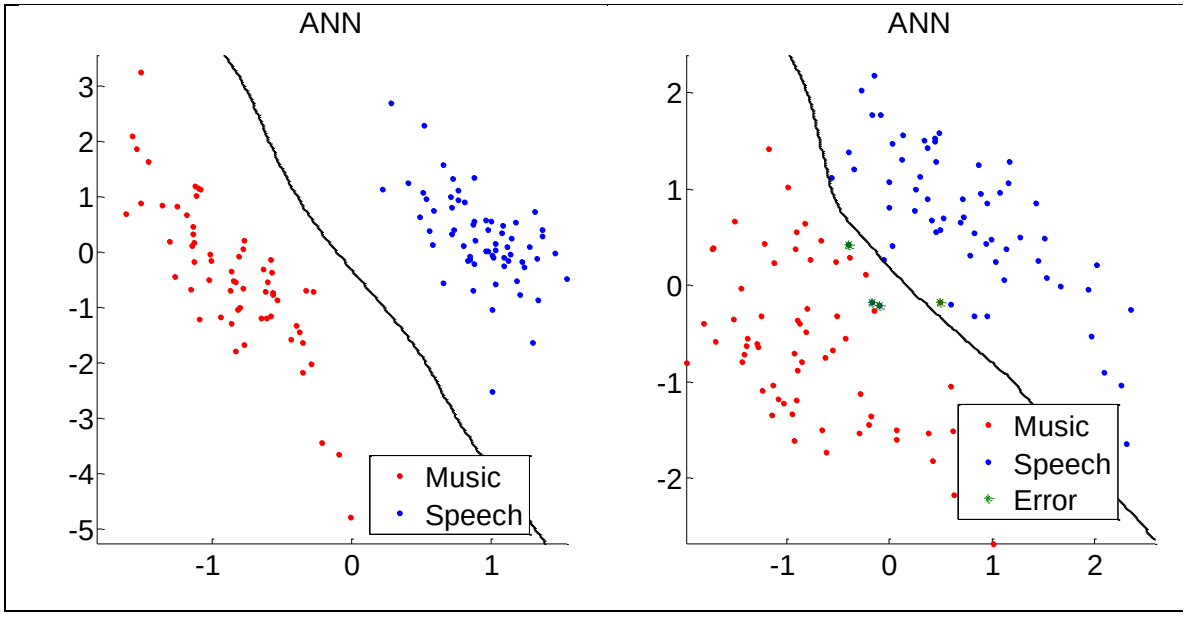
In the study case of the seven databases the results of an ANN classifier are shown in Table 10 and Figure 15, this network has 20 neurons in its hidden layer. Table 11 shows the results for a better number of neurons adapted for each database, the most interesting result here is the 100% accuracy in the M/S.P3 dataset, which is non-linearly separable.

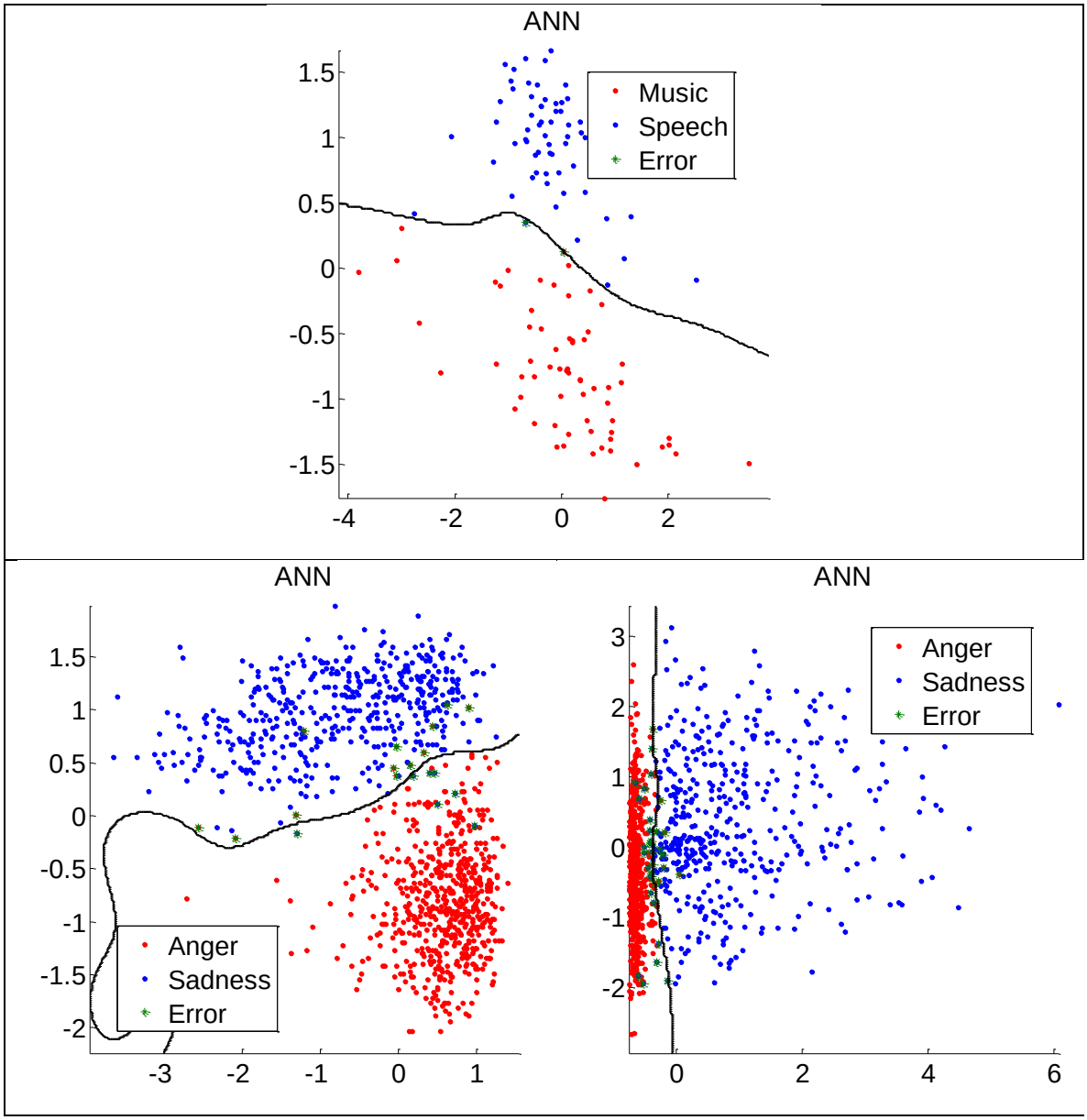
| Database | Accuracy |
|---|---|
| M/S.P1 | 100% |
| M/S.P2 | 96.9% |
| M/S.P3 | 98.4% |
| Br.1 | 98.2% |
| Br.2 | 96.9% |
| Br.3 | 80% |
| Br.4 | 99.4% |

**Table 10 ANN accuracy for each of the datasets**

| Database | Accuracy | Neurons |
|---|---|---|
| M/S.P1 | 100% | 5 |
| M/S.P2 | 99.2% | 70 |
| M/S.P3 | 100% | 22 |
| Br.1 | 98.4% | 30 |
| Br.2 | 97.2% | 22 |
| Br.3 | 80.3% | 29 |
| Br.4 | 99.3% | 25 |

**Table 11 Better approximation for the number of neurons for each dataset with an ANN and their accuracy**




ANN
Music
Speech
Error
ANN
Anger
Sadness
Error
ANN
Anger
Sadness
Error

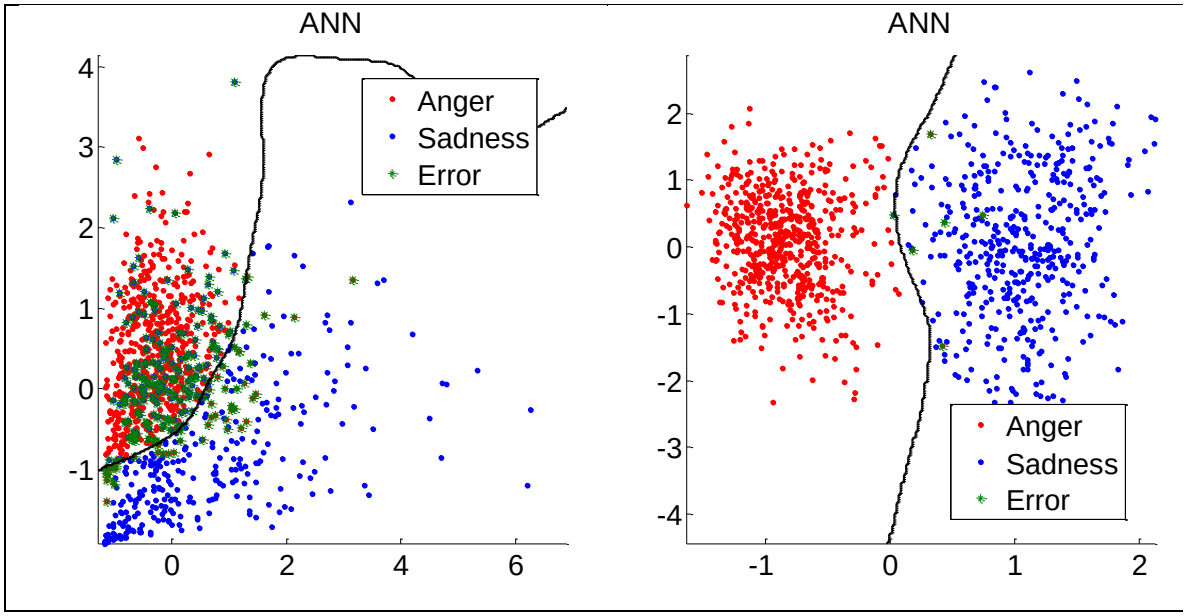



**Figure 15 ANN decision boundaries and classification results for M/S.P1, M/S.P2, M/S.P3, Br.1, Br.2, Br.3 and Br.4.**

## 5.4. Kernels

In the context of classification a kernel is a function $k(x, y)$ that substitutes the inner product $x^T y$, the result is equivalent to solving the same problem in a space with a higher dimensionality where the actual inner product is the one defined by the kernel function. The result of using a kernel is that a linear classification in the high dimensional space is equivalent to a nonlinear one in the original space (Herbrich, 2002).

The two kernel functions used in this work are the radial basis function (RBF):

$$k(x, y) = \exp(-\frac{\|x - y\|^2}{\sigma^2}) \qquad (12)$$

And the polynomial function (poly):

$$k(x, y) = (x^T y + \beta)^n \qquad (13)$$

A lot of classifiers can be modified with a kernel, this sections covers the results for a perceptron with both a RBF and poly kernel, and for a SVM with a RBF kernel.

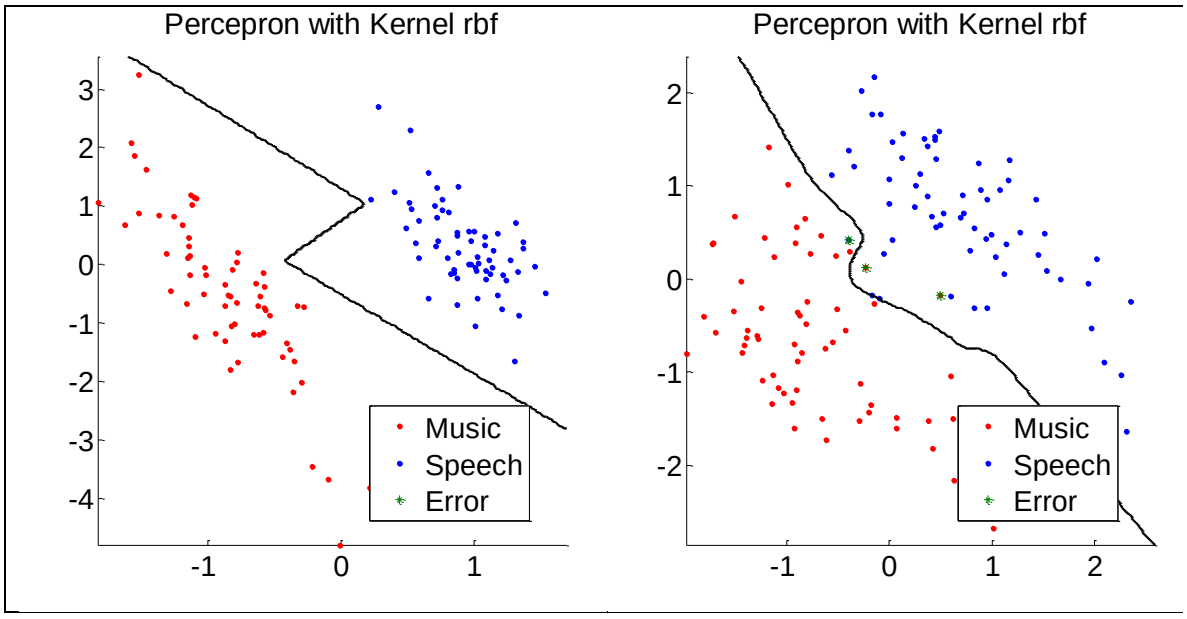
Table 12 and Figure 16 present the results in the databases for the perceptron with a RBF kernel with a parameter $\sigma = 0.2$. Table 13 shows a selection of a better parameter for each of the databases.

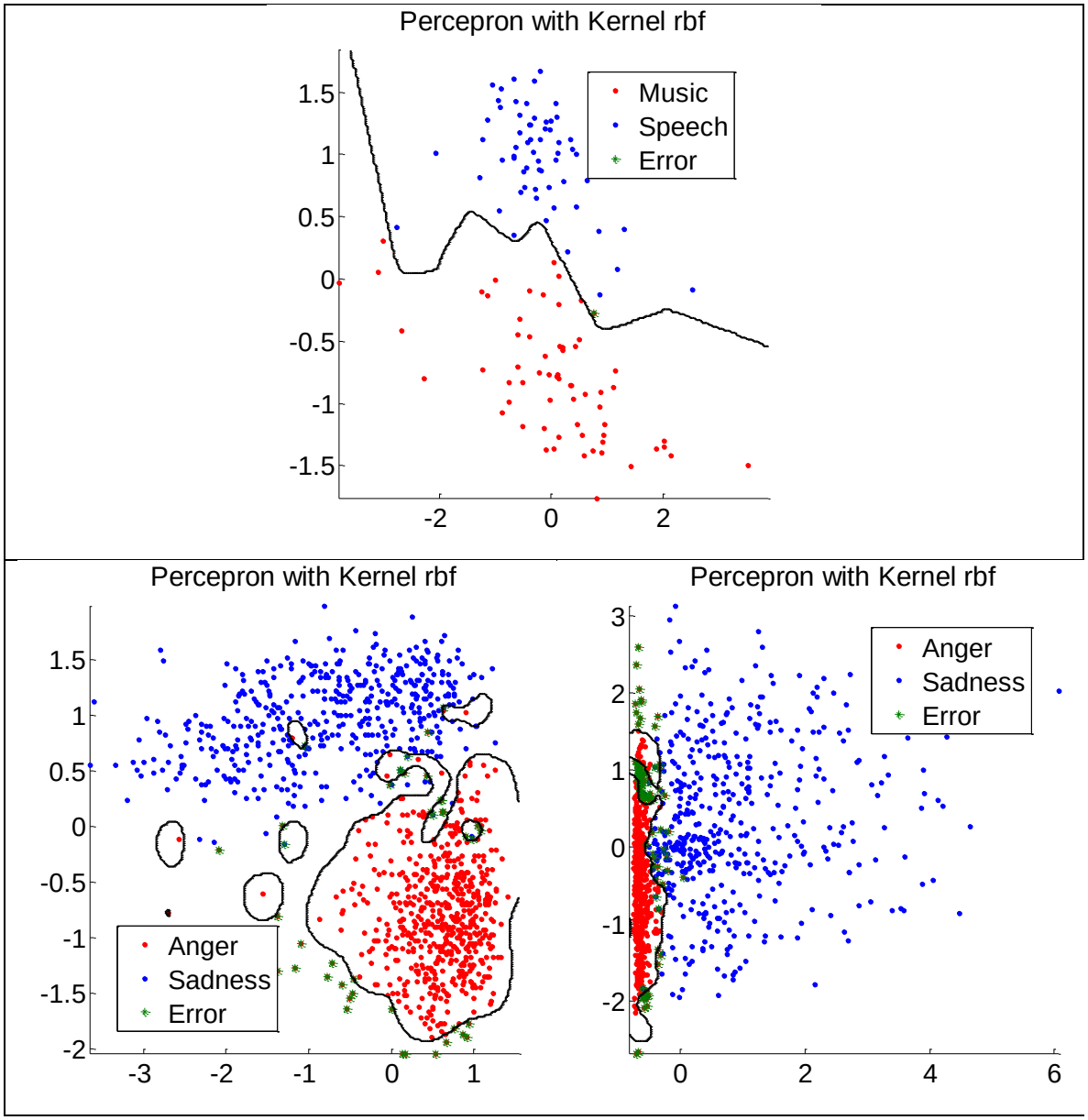
| Database | Accuracy |
|---|---|
| M/S.P1 | 99.2% |
| M/S.P2 | 97.7% |
| M/S.P3 | 99.2% |
| Br.1 | 96% |
| Br.2 | 91.8% |
| Br.3 | 54.9% |
| Br.4 | 99.3% |

**Table 12 Perceptron with RBF kernel accuracy for each dataset**

| Database | Accuracy | $\sigma$ |
|---|---|---|
| M/S.P1 | 100% | 0.1 |
| M/S.P2 | 97.7% | 0.2 |
| M/S.P3 | 100% | 0.1 |
| Br.1 | 96% | 0.2 |
| Br.2 | 92.6% | 0.3 |
| Br.3 | 89.6% | 0.01 |
| Br.4 | 99.3% | 0.2 |

**Table 13 Better parameter approximation for the RBF parameter for each dataset and the perceptron's accuracy**




Percepron with Kernel rbf
Music
Speech
Error
Percepron with Kernel rbf
Anger
Sadness
Error
Percepron with Kernel rbf
Anger
Sadness
Error

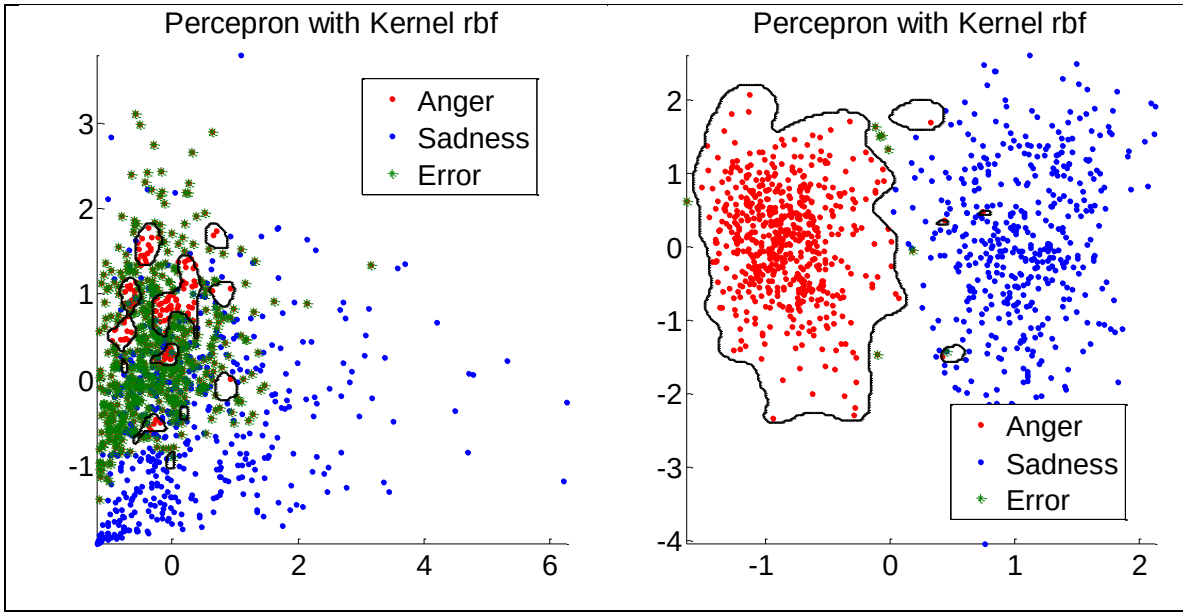



**Figure 16 Perceptron with RBF kernel decision boundaries and classification results for M/S.P1, M/S.P2, M/S.P3, Br.1, Br.2, Br.3 and Br.4.**

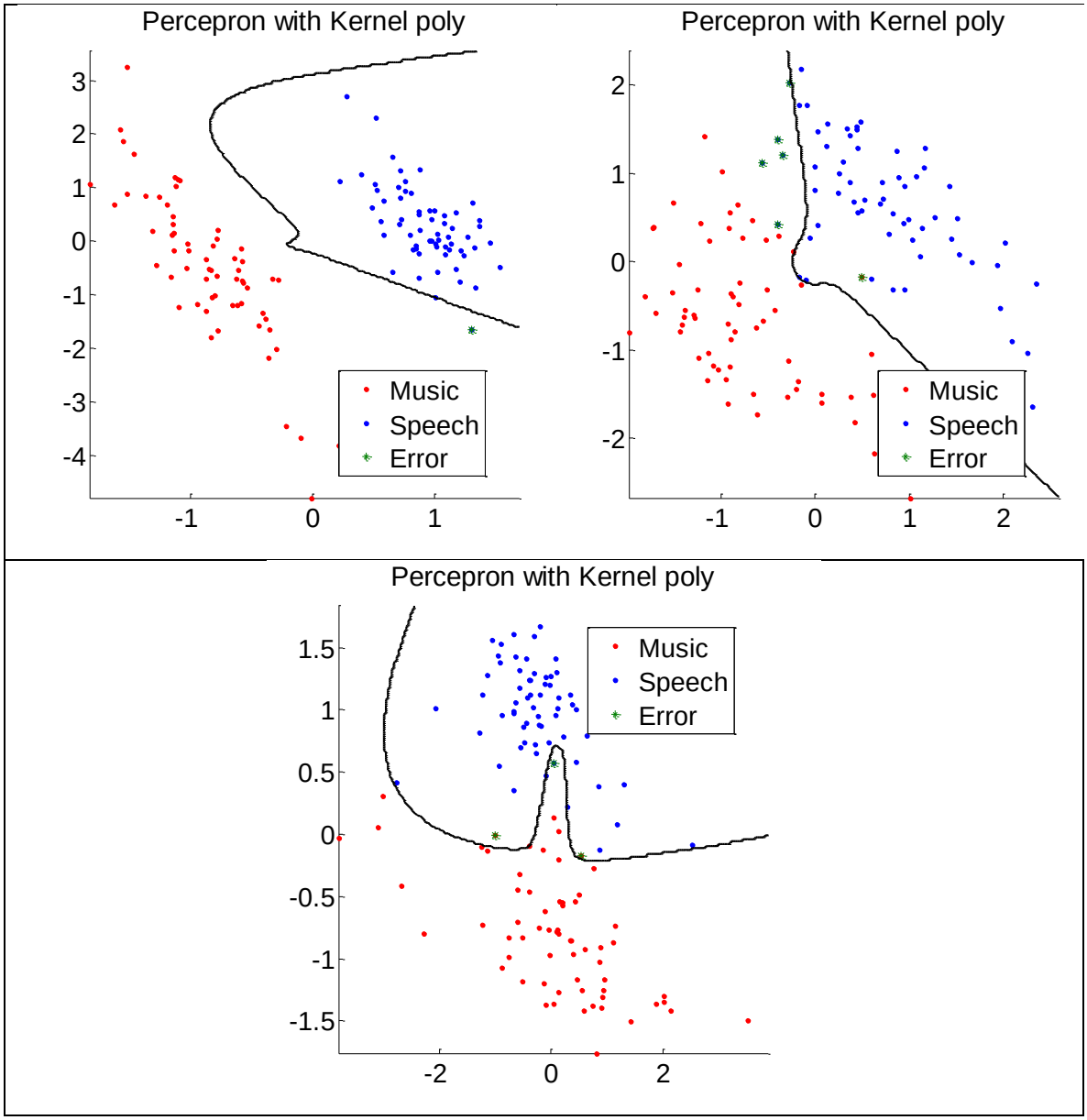
Table 14 and Figure 17 show the results of a perceptron with a polynomial kernel for the seven databases with parameters $\beta = 1$ and $n = 10$. Table 15 presents a better approximation of these parameters for each of the databases.

| Database | Accuracy |
|---|---|
| M/S.P1 | 98.4% |
| M/S.P2 | 95.3% |
| M/S.P3 | 97.7% |
| Br.1 | 97% |
| Br.2 | 95.1% |
| Br.3 | 71.1% |
| Br.4 | 98.9% |

**Table 14 Perceptron with poly kernel accuracy for each dataset**

| Database | Accuracy | $\beta$ | $n$ |
|---|---|---|---|
| M/S.P1 | 100% | 1 | 5 |
| M/S.P2 | 96.1% | 1 | 8 |
| M/S.P3 | 100% | 0.5 | 17 |
| Br.1 | 97% | 1 | 10 |
| Br.2 | 92.6% | 1 | 10 |
| Br.3 | 71.1% | 1 | 10 |
| Br.4 | 99.5% | 0.5 | 12 |

**Table 15 Better parameter approximation for the perceptron with poly kernel for each dataset and its accuracy**


Percepron with Kernel poly
Music
Speech
Error
Percepron with Kernel poly
Music
Speech
Error
Percepron with Kernel poly
Music
Speech
Error

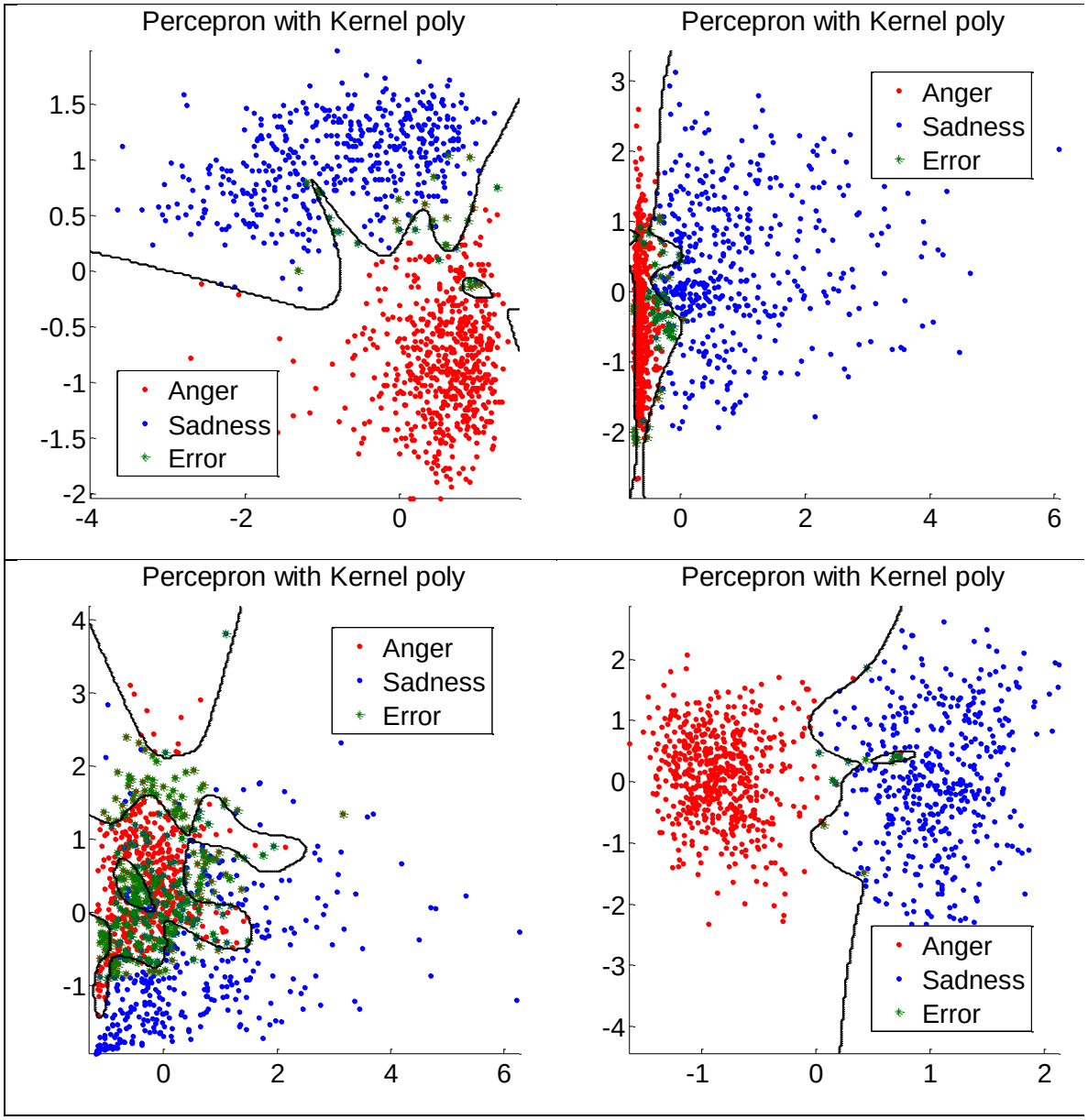



**Figure 17 Perceptron with poly kernel decision boundaries and classification results for M/S.P1, M/S.P2, M/S.P3, Br.1, Br.2, Br.3 and Br.4.**

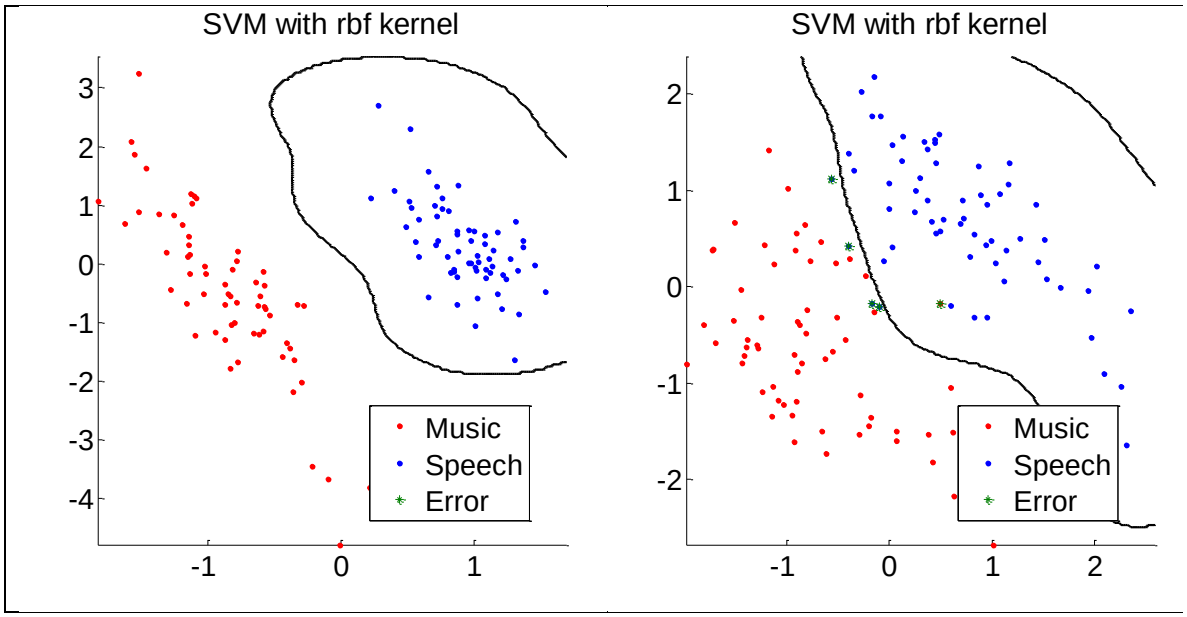
Finally the results for SVM with a RBF kernel with parameters $C = 0.5$ and $\sigma = 0.5$ for all databases are shown in Table 16 and Figure 18. Table 17 presents the results for better parameters for each database.

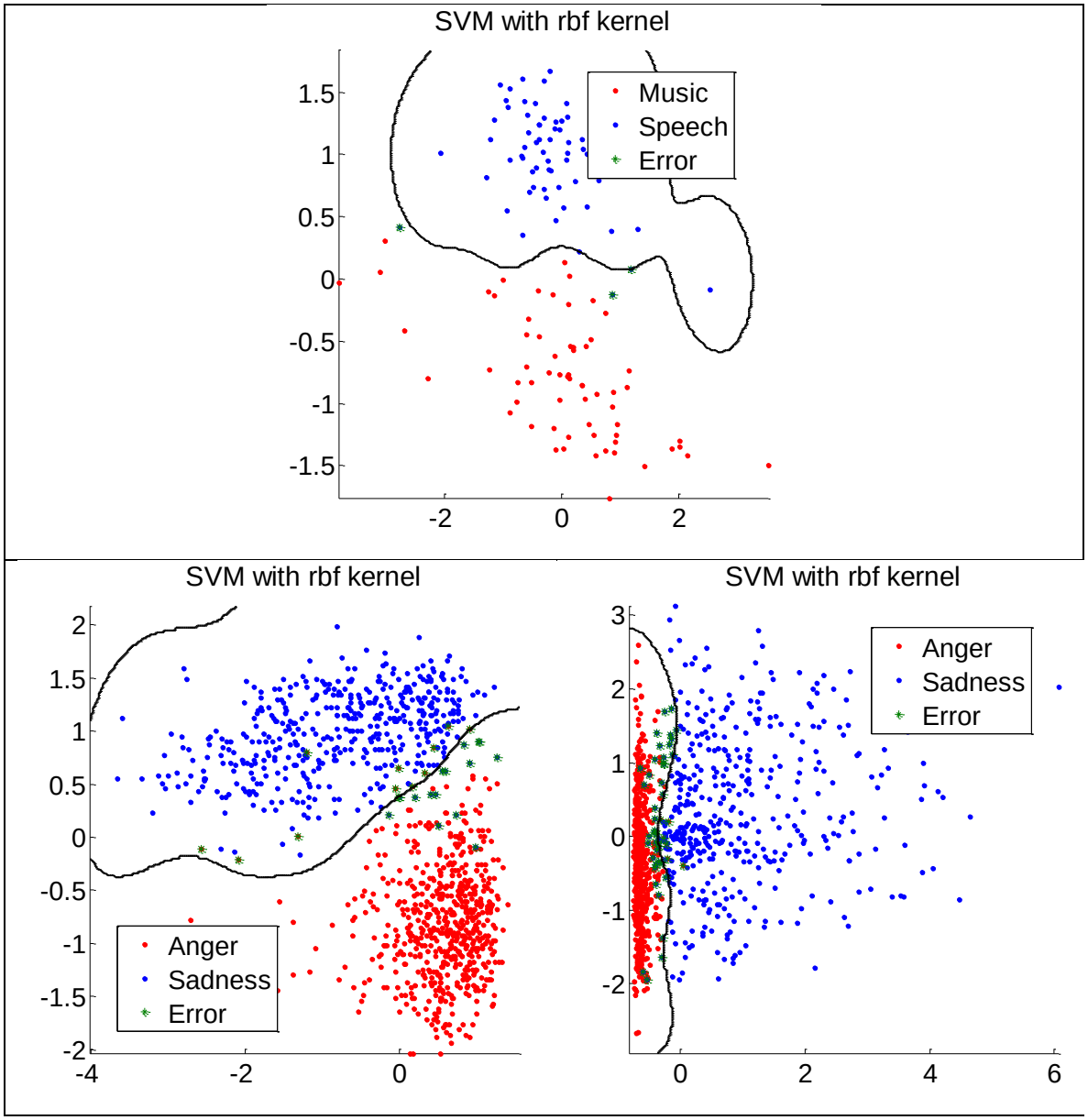
| Database | Accuracy |
|---|---|
| M/S.P1 | 99.2% |
| M/S.P2 | 96.1% |
| M/S.P3 | 97.7% |
| Br.1 | 97.5% |
| Br.2 | 95.6% |
| Br.3 | 56.3% |
| Br.4 | 99.2% |

**Table 16 Accuracy in each dataset for a SVM with RBF kernel**

| Database | Accuracy | $C$ | $\sigma$ |
|---|---|---|---|
| M/S.P1 | 100% | 0.5 | 5 |
| M/S.P2 | 96.1% | 0.2 | 0.5 |
| M/S.P3 | 99.2% | 0.3 | 1 |
| Br.1 | 97.9% | 0.3 | 0.01 |
| Br.2 | 96.6% | 0.08 | 10 |
| Br.3 | 78.4% | 0.51 | 0.5 |
| Br.4 | 99.2% | 0.5 | 0.5 |

**Table 17 Better parameters for each dataset and accuracy for the SVM with RBF kernel**




SVM with rbf kernel
Music
Speech
Error
SVM with rbf kernel
Anger
Sadness
Error
SVM with rbf kernel
Anger
Sadness
Error

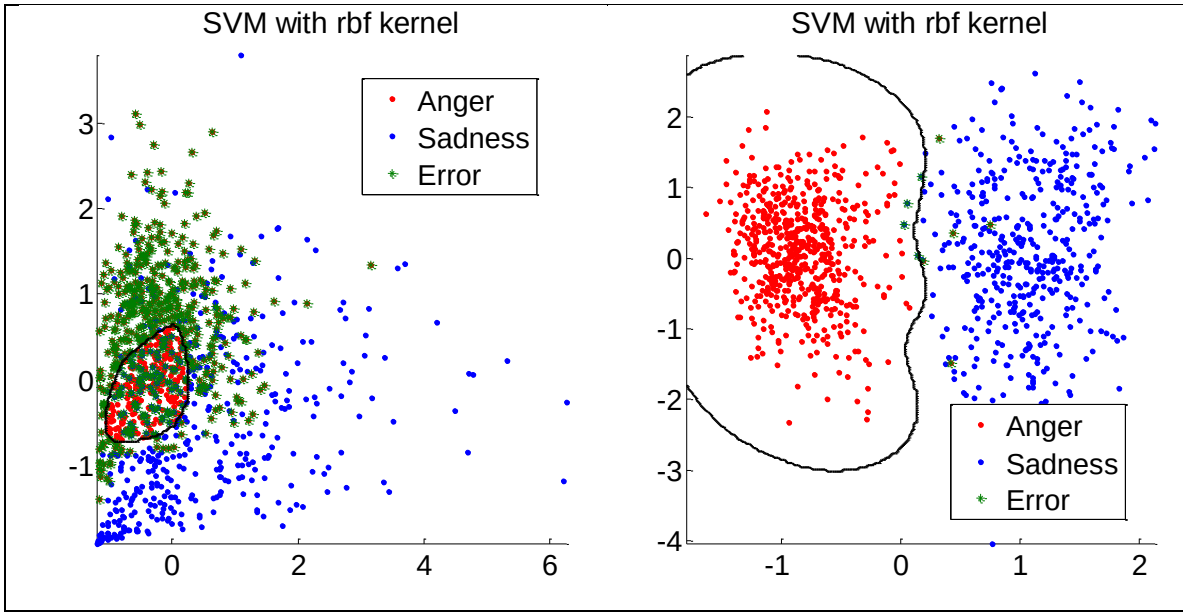



**Figure 18 SVM with RBF kernel decision boundaries and classification results for M/S.P1, M/S.P2, M/S.P3, Br.1, Br.2, Br.3 and Br.4.**

## 5.5. Classification results for datasets

This section presents the best result for each dataset and which classifier achieved it Table 18, it is not meant to be a guide for the selection of a classifier but to report in a compact manner the above mentioned result.

| Database | Accuracy | Classifier |
|---|---|---|
| M/S.P1 | 100% | Any |
| M/S.P2 | 99.2% | ANN |
| M/S.P3 | 100% | Perceptron with poly kernel<br>Perceptron with RBF kernel<br>ANN |
| Br.1 | 98.4% | ANN |
| Br.2 | 98% | Tree |
| Br.3 | 89.6% | Perceptron with RBF kernel |
| Br.4 | 99.6% | Tree |

**Table 18 Best accuracy for each dataset and its corresponding classifier after parameter tuning**

# 6. Emotion recognition from speech

This section provides a quick overview of emotion recognition from speech, how this problem was approached and the results obtained in the whole Berlin Database of Emotional Speech. The main idea is both to provide a full study case for classification as well as the best results obtained for this specific problem.

## 6.1. Background

Understanding the emotional state of a speaker is a great step towards a natural interaction between man and machine. Automatic speech emotion recognition is a recent research field, which is defined as extracting the emotional state of a speaker from his or her speech (El Ayadi, Kamel, & Karray, 2011).

The features extracted from the speech signals vary widely, usual features include pitch, formats, energy, timing, voice quality, spectral, etc. (El Ayadi, Kamel, & Karray, 2011). Some works (Kandali, Routray, & Basu, 2009) (Degaonkar & Apte, 2013) (Zhiyan & Jian, 2013) use the decomposition of the signal using wavelet transform for feature extraction.

The accuracy of classifiers in the automatic speech emotion recognition task is in average between 51.19% and 70% for ANN and between 74% and 81.9% for the other ones (HMM, GMM, SVM), (El Ayadi, Kamel, & Karray, 2011), also it is important to note that for speaker-independent speech emotion recognition systems the accuracy is less than 80% in most of the mentioned techniques, but for speaker-dependent classification, the recognition accuracy can exceeded 90% (El Ayadi, Kamel, & Karray, 2011). The classification methods employed include support vector machines (SVM), hidden markov models (HMM), Gaussian mixture models (GMM), artificial neural networks (ANN), k-NN, etc. (Phinyomark, Limsakul, & Phukpattaranont, 2009) (Ntalampiras & Fakotakis, 2012) (Schuller, Batliner, Steidl, & Seppi, 2011) (Lee & Narayanan, 2005) (El Ayadi, Kamel, & Karray, 2011)) (Fernandez & Picard, 2005) (Theodoridis & Koutroumbas, 2009).

## 6.2. Methodology and results

Using the whole database with its seven emotions (happiness, neutral, boredom, anxiety, sadness, disgust and anger) widowed with a rectangular window of 1s with an overlap between windows of 0.4s a total of 2710 windows.

### 6.2.1. Feature extraction

From each window a multiresolution analysis of 10 levels was performed (Mallat, 1989) with the Haar, Daubechies 6, 8 and 10 wavelets. In each level of decomposition the features presented in Table 1 were measured. Dynamic features inspired by (Ntalampiras & Fakotakis, 2012) were also measured, for this case the signal is re-windowed in windows of 30 ms with an overlap of 10 ms. Between the smaller windows, the Table 1 features are extracted from the multiresolution analysis.

Later, the dynamic features are easily calculated from the smaller windows over the 1-second original segment, these measurements are shown in Table 19. After cleaning the data from features which contained the same information as well as those that didn't contain any the total number of features is 11032.

| Number | Dynamic Feature |
|---|---|
| 1 | Mean |
| 2 | Standard deviation |
| 3 | Maximum |
| 4 | Minimum |
| 5 | Kurtosis |
| 6 | Statistical Asymmetry |
| 7 | Minimum of the absolute value |
| 8 | First coefficient of an order 2 AR |
| 9 | Second coefficient of an order 2 AR |

**Table 19 Dynamic features**

### 6.2.2. Feature selection and classification

From this set of features three different feature selection methods were performed:

In first place a sequential floating forward feature selection with a scatter matrix (Pudil, Novovičová, & Kittler, Floating search methods in feature selection, 1994). In this case two features for each possible pair of emotions were found, for a total of 42 features. After removing the repeated features the final set of 29 is shown in Table 20.

| Multiresolution analysis information | Feature measurement | Dynamic measure |
|---|---|---|
| - | 1 | - |
| - | 16 | - |
| Haar Level 10 of approximation | 20 | - |
| Haar level 7 of detail | 8 | - |
| Haar level 6of detail | 11 | - |
| Haar levl 6 detail | 14 | - |
| Haar level 2 of detail | 14 | - |
| Haar  level 4 of approximation | 10 | - |
| Db6 level 1 of detail | 8 | - |
| Db6 level 4 of approximation | 6 | - |
| Db8 level 9 of detail | 10 | - |
| Db8 level 9 of detail | 11 | - |
| Db8 level 2 of detail | 22 | - |
| Db8 level 1 of approximation | 1 | - |
| Db8 level 4 of approximation | 1 | - |
| Haar level 8 of approximation | 16 | 1 |
| Haar level 6 of detail | 23 | 3 |
| Haar level 7 of approximation | 11 | 5 |
| Haar level 1 of approximation | 21 | 6 |
| Haar level 1 of detail | 10 | 8 |
| Haar level 1 of approximation | 11 | 8 |
| Haar level 2 of approximation | 14 | 8 |
| Haar level 3 of approximation | 11 | 8 |
| Haar level 3 of approximation | 13 | 8 |
| Haar level 3 of approximation | 14 | 8 |
| Haar level 3 of approximation | 15 | 8 |
| Haar level 5 of approximation | 2 | 8 |
| Haar level 7 of approximation | 20 | 8 |
| Haar level 7 of detail | 21 | 9 |

**Table 20 Final 29 features selected with sequential floating forward feature selection.**

After selection 29 features are identified, with this an ANN was trained using Bayesian Regulation Backpropagation (MacKay, 1992) (Foresee & Hagan, 1997) with 70% of the data for training, 15%

for testing and 15% for validation. The result is shown in Figure 19 that corresponds to the confusion matrix, the overall accuracy being 94.3%.

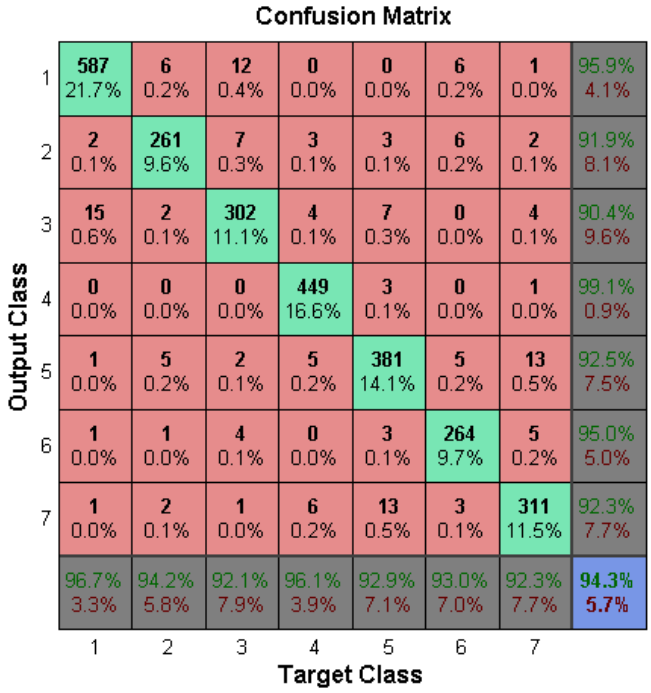



**Figure 19 Confusion matrix for the ANN with the sequential floating forward feature selection features.**

The second selection performed is an exhaustive search (Theodoridis & Koutroumbas, 2009), also in this case two features for each possible pair of emotions was found, for a total of 42 possible features. After removing the repeated features the final set consists of the 12 shown in Table 21.

| Multiresolution analysis information | Feature measurement | Dynamic measure |
|---|---|---|
| Haar level 7 of detail | 2 | 1 |
| Haar level 7 of detail | 21 | 1 |
| Haar level 7 of detail | 2 | 2 |
| Haar level 7 of detail | 21 | 2 |
| Haar level 7 of detail | 2 | 3 |
| Haar level 7 of detail | 21 | 3 |
| Haar level 8 of approximation | 15 | 4 |
| Haar level 7 of detail | 2 | 5 |
| Haar level 7 of detail | 21 | 5 |
| Haar level 7 of detail | 2 | 6 |
| Haar level 7 of detail | 21 | 6 |
| Haar level 8 of approximation | 1 | 7 |

**Table 21 Features selected with exhaustive search**

With this feature set an ANN was trained using Bayesian Regulation Backpropagation (MacKay, 1992) (Foresee & Hagan, 1997) with 70% of the data for training, 15% for testing and 15% for validation. The result is shown in Figure 20 which corresponds to the confusion matrix, the overall accuracy being 48.9%.

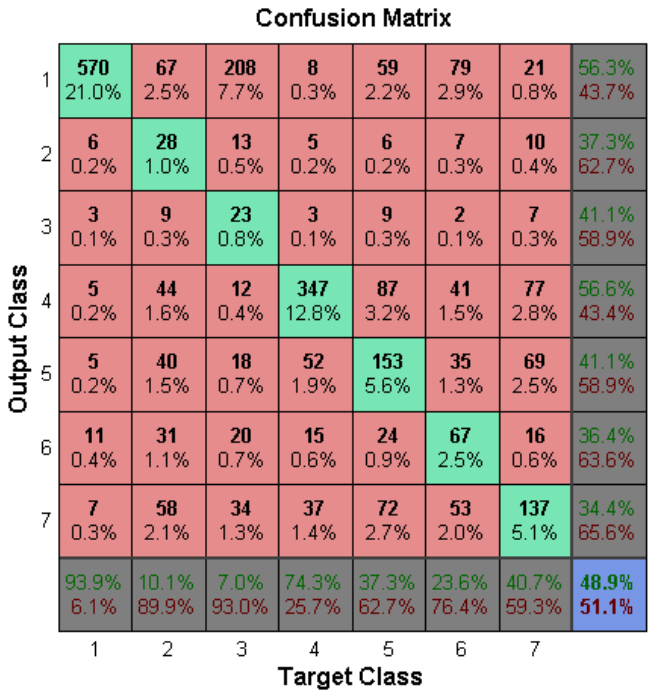



**Figure 20 Confusion matrix for the ANN with exhaustive search selected features**

The third and final selection performed with a genetic algorithm (Leardi, Boggia, & Terrile, 1992) with the classification accuracy of an LDA as the fitness criterion, 25 trails were performed. In each trail a maximum of 15 features of the set were selected, with 1000 sets as the population and a replacement rate of 50%, the mutation probability was 30% and each trail consisted of 50 generations. After all the trails the best feature set from the final generation was selected and is shown in Table 22.

| Multiresolution analysis information | Feature measurement | Dynamic measure |
|---|---|---|
| - | 1 | - |
| - | 6 | - |
| Haar level 10 of approximation | 11 | - |
| Haar level 5 of detail | 6 | - |
| Haar level 4 of detail | 7 | 1 |
| Haar level 3 of detail | 12 | 1 |

| | | |
|---|---|---|
| Haar level 5 of detail | 1 | 2 |
| Haar level 3 of detail | 5 | 3 |
| Db8 level 1 of detail | 20 | 1 |
| Db8 level 1 of detail | 10 | 4 |
| Db10 level 4 of approximation | 8 | 1 |
| Db10 level 4 of approximation | 11 | 1 |
| Db10 level 1 of detail | 6 | 1 |

**Table 22 Features selected with the genetic algorithm**

With this set of 13 features an ANN was trained using Bayesian Regulation Backpropagation (MacKay, 1992) (Foresee & Hagan, 1997) with 70% of the data for training, 15% for testing and 15% for validation. The result is shown in Figure 21 which corresponds to the confusion matrix, the overall accuracy being 84.8%.

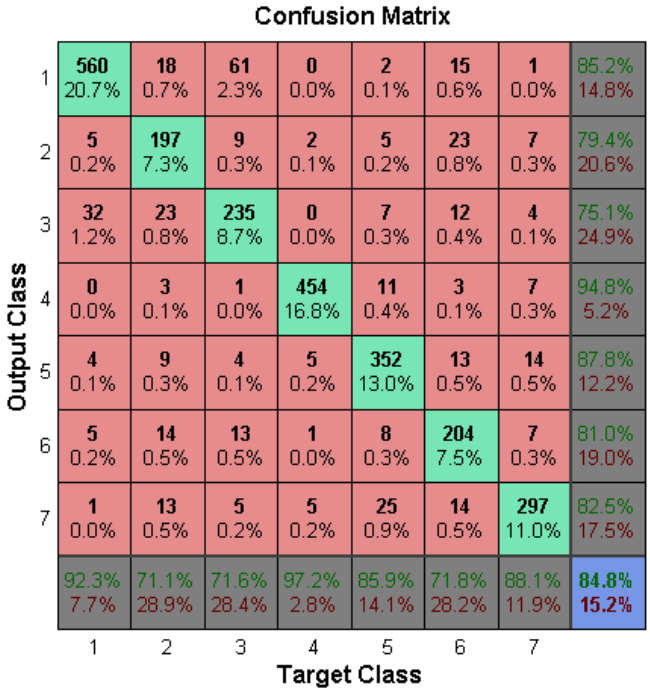



**Figure 21 Confusion matrix for an ANN with the features selected by the genetic algorithm**

# 7. Conclusions

An overview of the methodology for classification has been presented going through all the stages for both study cases. The best results for these are a 100% classification with the PCA projection of the selection of three features with any of the classifiers for the M/S database and between 99.2% and 99.6% for the Br.4 dataset giving a great differentiation between anger and sadness.

Not only the performance of each classifier presented was shown, but also how its decision boundary behaves with different datasets going from linearly separable passing through non-linearly separable all the way to non-separable classes. This in itself is really interesting and helpful in the understanding of a classification problem.

Finally for the classification of the whole Berlin database three different feature selections were used. The best result obtained for the genetic algorithm which selects a vector of features that yields a good classification for all the classes (84.8%) which competes with the best reported works (El Ayadi, Kamel, & Karray, 2011) and at the same time keeps the number of features low at 13. When looking only at the overall performance of the classifier the 29 features selected with the sequential floating forward algorithm yield a 94.3% classification rate which appears to beat other reported works. These results are both due to the use of a good selection of features as well as the features themselves which incorporate both a wavelet based decomposition as well as a wide range of features, both static and dynamic. Further analysis and validation is need for this specific case.

# 8. References


Arlot, S., & Celisse, A. (2010). A survey of cross-validation procedures for model selection. *Statistics surveys*, 40-79.

Burkhardt, F., Paeschke, P., Rolfes, M., Sendlmeier , W., & Weiss, B. (2005). A Database of German Emotional Speech. *Proceedings Interspeech.* Lisboa, Portugal.

Cook, G. T. (2015, 07 15). *GTZAN Music Speech*. Retrieved 2015, from GTZAN Music Speech: http://marsyasweb.appspot.com/download/data_sets/

Degaonkar, V. N., & Apte, S. D. (2013). Emotion modeling from speech signal based on wavelet packet transform. *International Journal of Speech Technology*, 1-5.

Duda, R. O., Hart, P. E., & Stork, D. G. (2012). *Pattern classification.* John Wiley & Sons.

El Ayadi, M., Kamel, M. S., & Karray, F. (2011). Survey on speech emotion recognition: Features, classification schemes, and databases. *Pattern Recognition*, 572-587.

Fawcett, T. (2006). An introduction to ROC analysis. *Pattern recognition letter*, 861-874.

Fernandez, R., & Picard, R. W. (2005). Classical and novel discriminant features for affect recognition from speech. *Interspeech*, 473-476.

Foody, G. M., Mathur, A., Sanchez-Hernandez, C., & Boyd, D. S. (2006). Training set size requirements for the classification of a specific class. *Remote Sensing of Environment* , 1-14.

Foody, G. M., McCulloch, M. B., & Yates, W. B. (1995). The effect of training set size and composition on artificial neural network classification. *International Journal of Remote Sensing*, 1707-1723.

Foresee, F. D., & Hagan, M. T. (1997). Gauss-Newton approximation to Bayesian learning. *Neural Networks, 1997., International Conference*, (pp. 1930-1935).

Friedman, J., Hastie, T., & Tibshirani, R. (2001). *The elements of statistical learning.* Berlin: Springer.

Giannakopoulos, T., & Pikrakis, A. (2014). *Introduction to Audio Analysis: A MATLAB® Approach.* Academic Press.

Herbrich, R. (2002). *Learning kernel classifiers.* Cambridge: MIT Press.

Hodge, V. J., & Austin, J. (2004). A survey of outlier detection methodologies. *Artificial Intelligence Review*, 85-126.

Ibrahim, R. K., Ambikairajah, E., Celler, B. G., & Lovell, N. H. (2007 ). Time-Frequency Based Features for Classification of Walking Patterns. *Digital Signal Processing.* Sydney: IEEE.

Jolliffe, I. (2002). *Principal component analysis.* John Wiley & Sons, Ltd.

Kandali, A. B., Routray, A., & Basu, T. K. (2009). Vocal emotion recognition in five native languages of Assam using new wavelet features. *International Journal of Speech Technology*, 1-13.

Karkanis, S., Iakovidis, D. K., Maroulis, D. E., Karras, D., & Tzivras, M. (2003). Computer-aided tumor detection in endoscopic video using color wavelet features. *Information Technology in Biomedicine* (pp. 141-152). IEEE.

Khan, S., & Alpaydin, E. (2004). *Introduction to Machine Learning.* MIT Press.

Kohavi, R. (1995). A study of cross-validation and bootstrap for accuracy estimation and model selection. *Ijcai*, 1137-1145.

Lambrou, T., Kudumakis, P., Speller, R., Sandler, M., & Linney, A. (1998). Classification of audio signals using statistical features on time and wavelet transform domains. *Acoustics, Speech and Signal Processing* (pp. 3621 - 3624). Seattle: IEEE.

Leardi, R., Boggia, R., & Terrile, M. (1992). Genetic algorithms as a strategy for feature selection. *Journal of chemometrics*, 267-281.

Lee, C. M., & Narayanan, S. S. (2005). Toward detecting emotions in spoken dialogs. *Speech and Audio Processing*, 293-303.

Lu, L., Zhang, H.-J., & Jiang, H. (2002). Content analysis for audio classification and segmentation. *Speech and Audio Processing*, 504-516.

MacKay, D. J. (1992). Bayesian interpolation. *Neural computation* , 415-447.

Mallat, S. G. (1989). A theory for multiresolution signal decomposition: the wavelet representation. *Pattern Analysis and Machine Intelligence, IEEE* , 674-693.

McKinney, M. F., & Breebaart, J. (2003). Features for audio and music classification. *ISMIR*, (pp. 151-158).

Mitchell, T. M. (2006). *The Discipline of Machine Learning.*

Ntalampiras, S., & Fakotakis, N. (2012). Modeling the temporal evolution of acoustic parameters for speech emotion recognition. *Affective Computing*, 116-125.

Ntalampiras, S., & Fakotakis, N. (2012). Modeling the Temporal Evolution of Acoustic Parameters for Speech Emotion Recognition. *IEEE TRANSACTIONS ON AFFECTIVE COMPUTING*, 116-125.

Park, C.-S., Choi, J.-H., Nah, S.-P., Jang, W., & Kim, D. Y. (2008). Automatic modulation recognition of digital signals using wavelet features and SVM. *Advanced Communication Technology, 2008. ICACT 2008. 10th International Conference* (pp. 387-390). IEEE.

Phinyomark, A., Limsakul, C., & Phukpattaranont, P. (2009). A novel feature extraction for robust EMG pattern recognition.

Pudil, P., Ferri, F. J., Novovicova, J., & Kittler, J. (1994). Floating search methods for feature selection with nonmonotonic criterion functions. *In Proceedings of the Twelveth International Conference on Pattern Recognition.*

Pudil, P., Novovičová, J., & Kittler, J. (1994). Floating search methods in feature selection. *Pattern recognition letters* , 1119-1125.

Rish, I. (2001). An empirical study of the naive Bayes classifier. *IJCAI 2001 workshop on empirical methods in artificial intelligence* (pp. 41-46). New York: IBM.

Rosenblatt, F. (1958). The perceptron: a probabilistic model for information storage and organization in the brain. *Psychological review*, 386-407.

Rosenblatt, F. (1961). *Principles of neurodynamics. perceptrons and the theory of brain mechanisms.* CORNELL AERONAUTICAL LAB INC BUFFALO NY.

Scheirer , E., & Slaney, M. (1997). Construction and eval- uation of a robust multifeatures. *IEEE Transactions on Acoustics, Speech, and Signal Process* (pp. 1331–1334). IEEE.

Schuller, B., Batliner, A., Steidl, S., & Seppi, D. (2011). Recognising realistic emotions and affect in speech: State of the art and lessons learnt from the first challenge. *Speech Communication* , 1062-1087.

Sun, Z., Bebis, G., & Miller, R. (2002). Quantized wavelet features and support vector machines for on-road vehicle detection. *Control, Automation, Robotics and Vision* (pp. 1641-1646). IEEE.

Theodoridis, S., & Koutroumbas, K. (2009). *Pattern Recognition.* Academic Press.

Tzanetakis, G., Essl, G., & Cook, P. (2001). Audio analysis using the discrete wavelet transform. *Proc. Conf. in Acoustics and Music Theory Applications.*

van Rijsbergen, C. (1979). *Information retrieval.* London: Buttweworths.

Vapnik, V. (2013). *The nature of statistical learning theory.* Springer Science & Business Media.

Ververidis, D., & Kotropoulos, C. (2006). Emotional speech recognition: Resources, features, and methods. *Speech communication* , 1162-1181.

Welling, M. (2005). *Fisher linear discriminant analysis.* Department of Computer Science, University of Toronto .

Widrow, B. (1960). Adaptive switching circuits. *1960 WESCON convention record* (pp. 96-104). Institute of Radio Engineers.

Yang, H., Van Vuuren, S., & Sharma, S. (2000). Relevance of Time-frequency Features for Phonetic and Speaker-channel Classification. *Speech Communication* (pp. 35-50). Amsterdam: Elsevier Science.

Zhiyan, H., & Jian, W. (2013). Speech emotion recognition based on wavelet transform and improved HMM. *Control and Decision Conference (CCDC)* (pp. 3156-3159). IEEE.

Zongker, D., & Jain, A. (1996). Algorithms for feature selection: An evaluation. *Pattern Recognition*, (pp. 18-22).