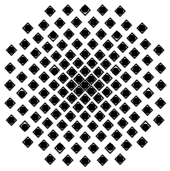

Universität Stuttgart
Institut für Signalverarbeitung und Systemtheorie
Prof. Dr.-Ing. B. Yang

PÜL
"Statistical Signal Processing —Pattern Recognition—"

Part I: Prostate cancer segmentation based on MRI and PET images
Part II: Speeker recognition based on the TIMIT database

Winterterm 2014/2015
23rd September 2014

Acknowledgements

We owe special thanks to:

Dr. med. Sergios Gatidis
Dr. rer. nat. Holger Schmidt
and Prof. Dr. med. Nina Schwenzer

from the PET/MR center of the Department of Diagnostic and Interventional Radiology at the University Hospital of Tübingen who kindly provided us with the medical data sets as well as with the results of their research and therefore laid the basis for this PÜL.

Contents

1. Introduction

1.1. Pattern recognition

Like discussed in the lecture “Detection and pattern recognition” [1], pattern recognition is a topic in machine learning covering a wide variety of problems. Methods of pattern recognition are commonly used in applications like the optical character recognition (OCR), facial recognition for security systems, scene analysis for driver assistance systems, spam detection or speech recognition systems.

This PÜL (Practical laboratory course) deals with the task of classification. The goal of classification is to assign class labels to objects of interest, whereat all objects assigned with the same class label shall share the same common properties. An example is the classification of vessels. Possible classes of vessels are for example ships and cars. In this case, a property shared by objects of the class car is that they have at least three wheels. Objects of the class ship share the property that they can displace a bigger mass of water than their own mass. Otherwise they would sink.

As depicted in Figure 1.1, classification consists of four basic processing steps. The process of classification can be devided into four basic processing steps. At first, observations are taken from each object using sensors. After preprocessing of the sensor raw data, meaningful features are extracted which can describe class-specific properties. The choice of an appropriate set of features greatly depends on the specific classification problem. The last step, classification means to estimate and assign class labels to each object. The basic idea is to find a decision boundary which separates the classes as good as possible in the feature space.

For the case of two classes and two features, this is visualized in Figure 1.2. Each point corresponds to an instance of a two-dimensional feature vector. One possible decision boundary is shown in black. The problem of classification is how to find the decision boundary which separates all possible instances of those two classes best without making this boundary too complex.

A variety of different classification techniques are known For detailed information about the most common techniques in classification you are encouraged to read [1], [2], [3], [4] or [5]

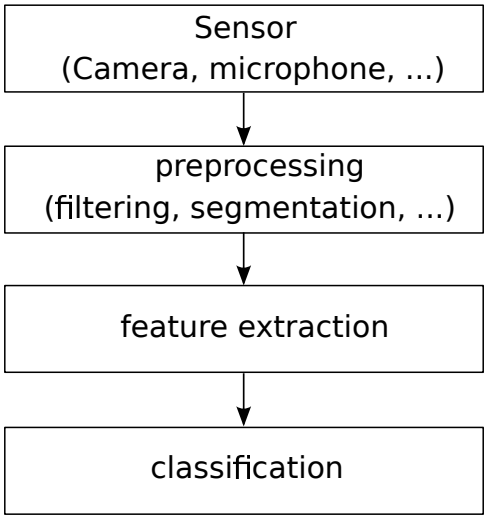

Figure 1.1.: Processing steps of classification

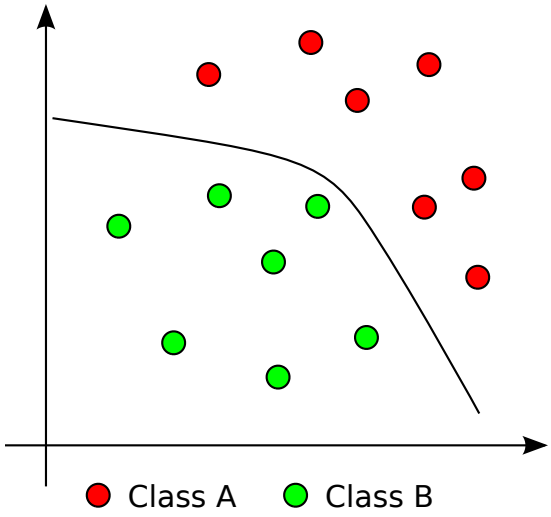

Figure 1.2.: Example of a binary classification problem in a two-dimensional feature space

1.2. Task overview

Part I: Prostate cancer segmentation

In the recent years, classification algorithms gained more and more influence in the area of medical signal processing. Classification algorithms are for example used for automatic compilation of clinical evidence. This means to automatically decide whether a patient has a specific disease or not. Those decisions are mostly based on features derived from measurements about the metabolism or the physiology of patients. This PÜL deals with automatic prostate cancer detection, what is a cutting edge problem from the area of medical signal processing [6].

At the Univeristy Hospital of Tübingen (UKT), 14 data sets from 14 different patients have been acquired for a study of prostate cancer detection. All of the 14 patients of this study suffer from prostate cancer at various stages and with different harshness. From each of them,

three dimensional images of the prostate have been obtained, using a magnetic resonance tomograph (MRT) as well as a positron emission tomograph (PET).

Up to today, detection of cancer tissue is done manually by experts who take a close look at the PET and MRT images of the patient and manually separate regions of healthy tissue, pixel-for-pixel, from those of prostate cancer. It is a decision which is called delineation of cancer tissue in medicine. The problem is that this manual delineation is inaccurate and time consuming. To speed up the process and to assure reproducible results, the objective is to use supervised classification algorithms to automatically separate those regions.

In terms of pattern recognition, this is a so called segmentation task. One possible result is shown in Figure 1.3. It is an MRT image of a slice of the prostate. Regions of healthy tissue are colored blue, whereas the regions of cancer are colored red. The non-colored region does not contain the prostate. More detailed information about the medical data set and related tasks are given in Chapter 2.

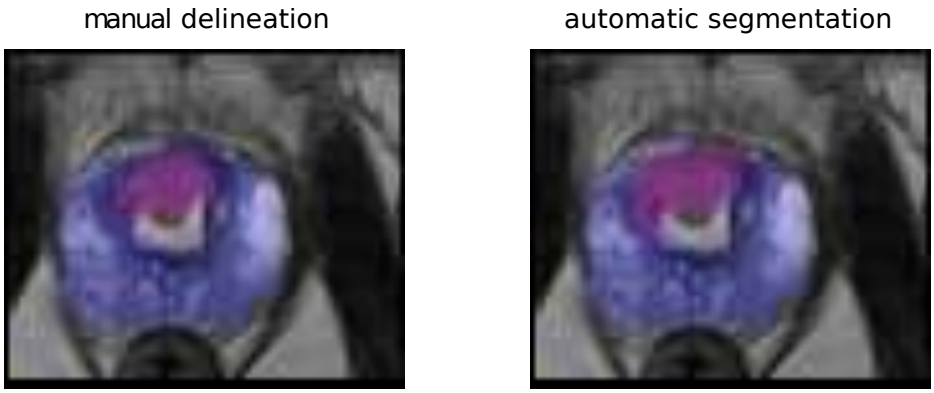

Figure 1.3.: Result of manual delineation and automatic segmentation

Part II: Speeker recognition

The identification and verification of speakers became relevant in several fields of application such as telephone banking, hotlines or surveillance in the last few years. The voice of a person is unique due to individual physical dimensions of the vocal tract, adapted speaking manners and because of differences in the used vocabulary. Therefore, speech signals can be used for speaker identification.

Very often, an easy approach for speaker identification is chosen which is based on the voice spectrum. Such an approach has been proposed in [] and acts as the basis of this part of the PÜL. Like depicted in Figure 1.4, for each time instance the aim is to estimate the speaker which has most likely spoken.

In this part of the PÜL, a simplified form of such a system shall be implemented. The decision shall be made sentence based, where each sentence contains only utterances from one speaker. Therefore, no sequence of different speakers has to be analyzed. From each sentence, different features have to be computed to create statistical speaker models. Those models can then be used to make a decision, which speaker has most likely spoken at which time instance.

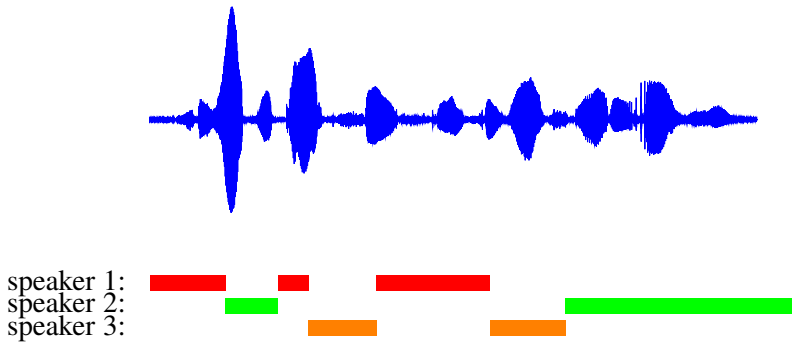

Figure 1.4.: Detection of multiple speakers in an audio material.

The features will be based on the spectral properties of the speech signals. In particular, a **Mel Filter Bank**, which consists of a number of triangular shaped bandpass filters, is applied. The bandwidth of the individual filters is defined by the logarithmic shaped **Mel Scale**, therefore the name Mel Filter Bank.

For each speaker, the spectral properties, then, are modeled by a **Gaussian Mixture Modell (GMM)**. The so obtained speaker models, then, are used to get a maximum likelihood estimate of which speaker has most likely spoken in the corresponding audio sample.

The audio samples to derive the speaker models and to test the recognition performance will be taken from the **TIMIT** database. Additionally, every team is encouraged to gather their own speech signals in order to test the performance of the implemented software.

Educational objective

The educational objective of this PÜL lies less on classification algorithms. Only easy classifiers shall be implemented from scratch by yourself. The implementation and debugging of advanced classification algorithms is very time consuming. Moreover, for most advanced classifiers there are a lot of off-the-shelf implementations available which have been extensively tested.

The main objectives of this PÜL are issues which you have not learned in the DPR lecture but which are still important for practical applications:

- The implementation of simple classifiers
- Methods of feature extraction
- Methods of feature normalization
- Feature selection to find subsets of meaningful features
- Choosing a classifier which is well suited for the given problem
- Application of training methods to find the best parameters. Overfitting has to be avoided in order to obtain low generalization errors
- Dealing with faulty training data. The given ground truth (class labels) may contain errors. Therefore, robust training techniques are needed (outlier detection/removal)

- The appropriate visualization of classification results.

1.3. Organisation of the lab course

During the PÜL the work will be carried out in teams of two students. The focus lies on self-dependent working. Problems are to be solved using informations from literature.

This script shall not act as a manual which contains all necessary steps, ultimately leading to a working solution of the posed problem. It shall rather act as a basic introduction into the topic, containing all theoretical background needed to understand the task. Furthermore, it shall sensitize every participant to the most important problems he will be dealing with.

In each group, the tasks should be split equally and reasonably in order to make fast progress. This means, especially that it should be avoided that team mates do the same implementation twice. The work may be carried out at home or in the institute laboratory. Nevertheless, every team has to be **present in the laboratory every week** during the semester.

At the end of the PÜL, each team has to write a **report** with **at least 30 pages**. The topics have to be split in a way that every student contributes to the report with at least 15 pages. The report should focus on problems which have been encountered during the lab course. Furthermore, the performance of the implemented algorithms should be assessed and distinct results should be discussed. As an official conclusion of the PÜL, there will be a colloquium where each team will present his results in a **speech of 15 minutes**.

2. Prostate cancer segmentation based on MRT and PET images

2.1. Motivation

Since people tend to become older in the recent centuries, prostate cancer has evolved to a widely spread and lethal disease effecting elderly people. As its name implies, prostate cancer is a type of cancer affecting the prostate, which is a gland of the size and shape of a conker in the male reproductive system. The exact location of the prostate is shown in Figure 2.1.

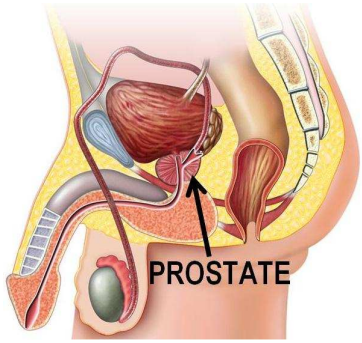

Figure 2.1.: Position of the prostate

Prostate cancer is one of the most frequently diagnosed cancer type for males. In fact, it is the leading cause of cancer death among men in the USA. In Germany, 3% of men die of prostate cancer [7]. Because of advances in medicine today, the diagnosis of prostate cancer does not mean immediate death for the patient. Knowing the type of the cancer as well as the regions which are affected, there are good ways to deal with this disease.

If prostate cancer is diagnosed, the first step is the localization of the ill tissue and the assessment of the damage. MRI and PET are promising non-invasive imaging techniques which are widely used for this localization step. Combining measurements of both to form multispectral images, specialists are able to manually delineate the tissue affected by the cancer.

However, manual delineation is inaccurate because the task involves a great amount of experience. The results from different experts are not comparable or reproducible because the assessment is subjective. Furthermore, manual delineation is very expensive because it is very time consuming and only well paid experts can be trusted with this task. Therefore, the aim is the automation of the delineation of cancer tissue in order to get more accurate and comparable results and in order to save cost.

2.2. The medical data set

The medical data set for automatic prostate cancer segmentation consists of five different 3D images derived from Positron Emission Tomography (PET) and from Magnetic Resonance Tomography (MRT). As shown in Figure 2.2, each of those images represents a cubic around the prostate of the patient. 2D images at a fixed z-coordinate are referred to as slices.

For each patient k, the dimensions $\underline{d}_k = (l_k, m_k, n_k)$ of all five 3D images are equal. However, due to variation of the prostate size among different patients, the dimensions of the images from different patiens are not the same and change between 34x41x31 and 320x320x43 pixels per image. In each image, approximately 35.000 voxels lie inside the prostate of the patient. In Figure 2.2, this region is marked gray. The rest of the voxels cover the surrounding tissue.

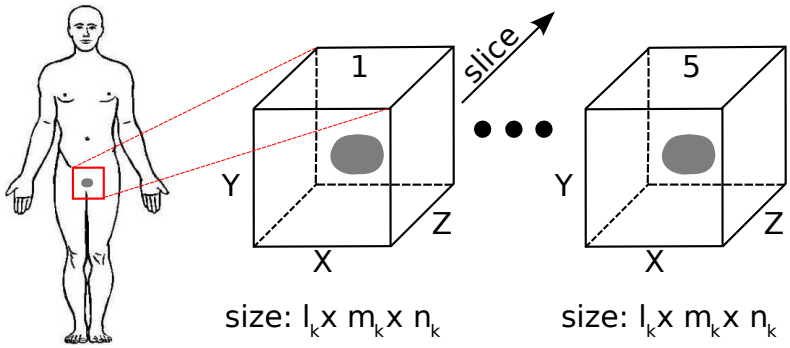

Figure 2.2.: Coordinate system of the medical data set

Example slices of the five 3D images of one patient are shown in Figure 2.3. They are the:

- T2 weighted MR image
- Apparent diffusion coefficient (ADC) map
- K_{trans} map
- K_{ep} map
- PET map

The left most four images are acquired with MRT, whereas the right most image is acquired with PET. The prostate is visible in the left most image. It is the conker shaped tissue covering most of the lower half of the image and is surrounded by the blue line. Regions where the cancer has spread are surrounded with a red line.

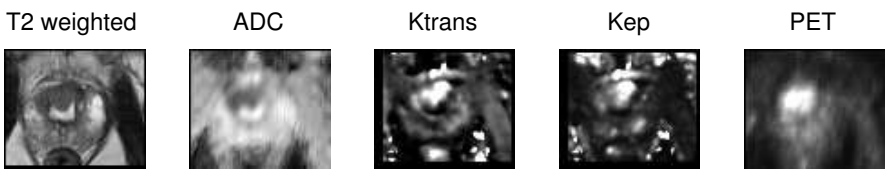

Figure 2.3.: Example data set for prostate cancer segmentation

T2 weighted MR images

In T2 weighted MR images, the structure and composition of the tissue is visualized. During measurement, the effect of nuclear magnetic resonance is used, i.e. nuclei in the tissue begin to radiate electromagnetic energy if they are excited properly. Thereby, the patient body is devided into small volume elements, called voxels. The amount of energy radiated from each of those voxels is measured and used to produce a 3D image of the patient.

Tissue which is rich of water appears bright in MR images, tissue with less water as well as fat or bone appear dark. Because cellular density is high in cancer tissue, there is less space left for water. Therefore, the signal intensity of cancer tissue is also lower if compared with healthy tissue. This can be seen in the leftmost image of Figure 2.3 where cancer tissue appears darker than the surrounding healthy tissue of the prostate.

The distribution of the voxel intensity is shown in Figure 2.4. The two peaks at the intensity $2e^4$ and $3.5e^4$ represent the cancer and non-cancer tissue. However, a separation based on the T2 weighted image only seems inpractical because the distributions of the intensity for cancer and non-cancer show a significant overlap.

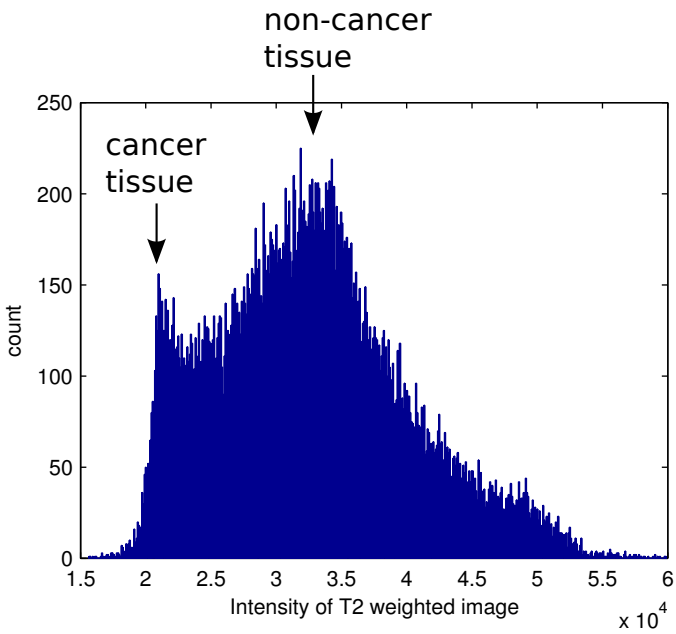

Figure 2.4.: Histogram of signal intensity of a T2 weighted MR image

Furthermore, a comparison of T2 weighted images from different measurements is difficult because those images are relative measures. The mean signal strength depends on the energy that can be used for excitation. It is limited by the amount of energy the body of the patient can dissipate as heat without causing thermal damage to the tissue. Therefore, distributions of voxel intensity from different measurements are shifted versions of the distribution shown in Figure 2.4.

For participants interested in MRT, further information about this imaging technique are given in the appendix.

The apparent diffusion coefficient (ADC)

The apparent diffusion coefficient (ADC) map measures the rate of diffusion of liquids through the tissue. As shown in Figure 2.5, for each point the ADC describes how fast liquids are able to move in three spatial directions. Those three rates of diffusion are named $D_x(x,y,z)$, $D_y(x,y,z)$ and $D_z(x,y,z)$. The absolute rate of diffusion is

$$D(x,y,z) = \sqrt{D_x^2 + D_y^2 + D_z^2}. \tag{2.1}$$

$D(x,y,z)$ is large if the liquids move fast and small if they move slowly.

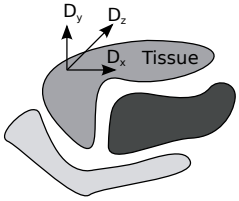

Figure 2.5.: Three-dimensional diffusion coefficient

Analysis of the data set shows that the rate of diffusion through tumour tissue is lower than in healthy tissue. This can be seen in the second image from the left in Figure 2.3. The tumour appears dark, whereas the healthy prostate tissue is bright. This is again because cells are more densely packed in tumor regions and therefore, liquids can not move as freely as in healthy tissue. The distribution of the ADC value for cancer and non-cancer tissue is shown in the histogram 2.6.

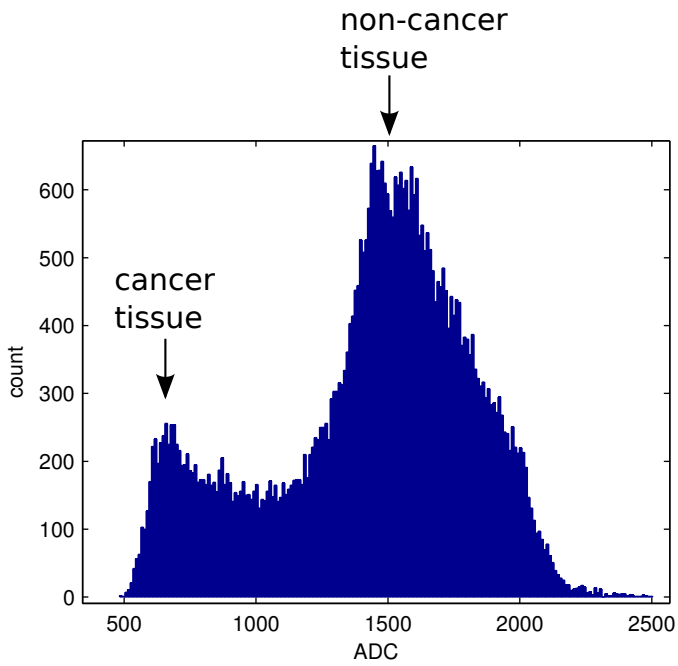

Figure 2.6.: Example histogram of ADC value from one patient

The mean value of the ADC for healthy tissue varies from patient to patient and ranges from $1300 < \mu_{ADC} < 1600$. This corresponds to the huge maximum at the right in Figure 2.6.

Variations occur, because the normal rate of diffusion in healthy tissue depends upon the patients electrolyte metabolism and is thus subject to variations.

The mean value for the ADC of cancer tissue is by a factor c lower with $0.4 < c < 0.7$. This corresponds to the small maximum at the left in Figure 2.6.

Diffusion parameters K_{trans} and K_{ep}

The diffusion parameters K_{trans} and K_{ep} also describe diffusion rates in the tissue. However, they have to be distinguished from the ADC map. As described in detail in [8] and [9], K_{trans} and K_{ep} are based on a compartment model of the human body and describe how chemical substances like drugs or contrast agents are brought into and out of those compartments. An example of the used two-compartment model is given in Figure 2.7.

The first compartment is the extravascular extracellular space (EES) which has the fractional volume v_e. It is body liquid outside any blood vessels in the space between the cells. The second compartment is the intravascular extracellular space (IES) which covers the blood plasma in the blood vessels and occupies the fractional volume v_p. Body fluids pass between those two compartments taking drugs and electrolytes with them.

The two diffusion coefficients describe the rate of diffusion between those compartments. K_{trans} describes how fast the tracer moves out of the blood plasma (IES) into the tissue (EES). K_{ep} is the rate of diffusion into the opposite direction and therefore describes how fast the tracer is flushed out of the tissue. The rate of diffusion between the compartments depends upon the concentration ratio of the contrast agent in the two compartments. The diffusion rates are calculated by solving the differential equation

$$\frac{\mathrm{d}C_t(t)}{\mathrm{d}t} - v_p \frac{\mathrm{d}C_p(t)}{\mathrm{d}t} = K_{trans} C_p(t) - k_{ep}(C_t(t) - v_p C_p(t)). \tag{2.2}$$

$C_p(t)$ is the concentration of the tracer in the plasma and $C_t(t)$ is the concentration of the tracer in the blood plasma. The higher the concentration $C_p(t)$, the faster will the tracer move into the tissue. On the other hand, if the difference $C_t(t) - v_p C_p(t)$ is big, the tracer will move fast out of the tissue back into the blood plasma.

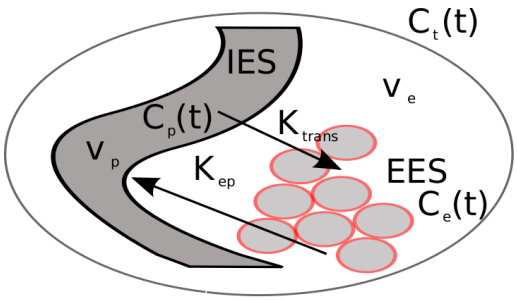

Figure 2.7.: Compartment model for diffusion parameters

Studies have shown that both diffusion parameters are higher in cancer tissue. Hence, the k_{trans} and the k_{ep} map are bright in regions of cancer tissue. This can be seen in the third and fourth image from the left in Figure 2.3. However, as shown in the histograms in Figure 2.8 and 2.9, the two classes are not separable.

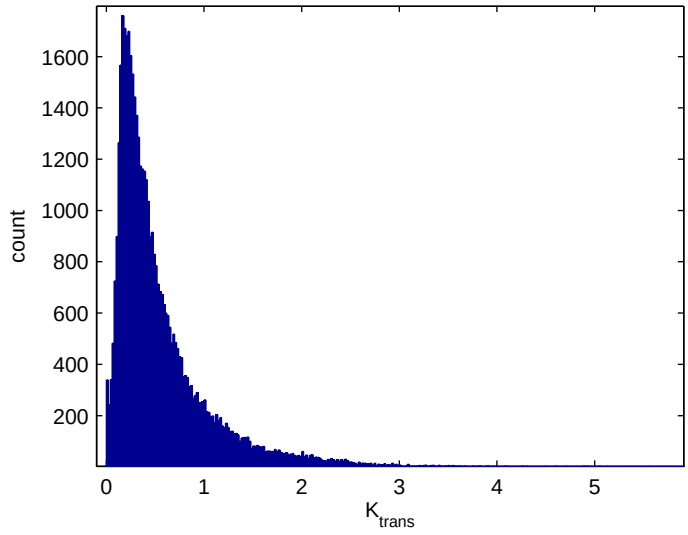

Figure 2.8.: Example histogram of diffusion koefficient K_{trans}

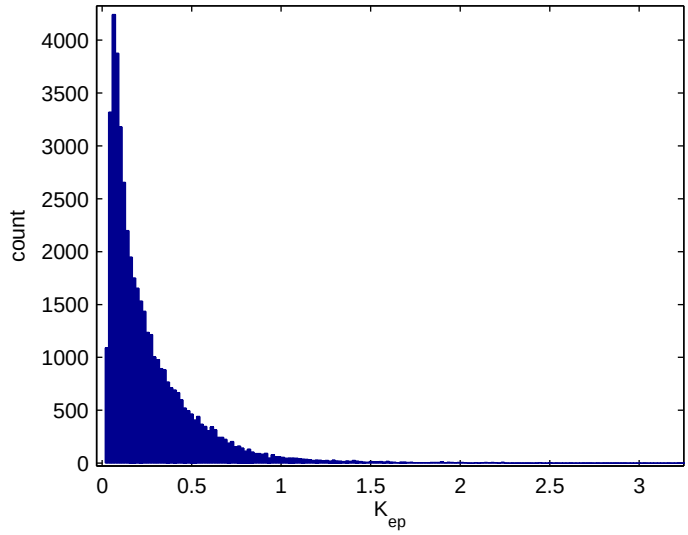

Figure 2.9.: Example histogram of diffusion koefficient K_{ep}

Positron Emission Tomography (PET) map

Positron emission tomography is an imaging technique which gives a 3D map of the concentration of radio nucleides in the human body and is described in detail in [10]. Before measurement, a radioactive substance, the so called tracer, is injected into the patients body. Most oftenly flourodeoxyclucose is used for this purpose. It is a type of sugar which is utilized by cells of the tissue just as normal glucose.

After injection the tracer spreads in the body of the patient and the cells start to accumulate the tracer through their normal metabolism. After a while the patient is placed in the PET scanner which measures the local radioactivity of the patients tissue. This measured radioactivity is known as the Standardized Uptake Value (SUV). The SUV describes how much

sugar has been used in which regions due to different rates of metabolism. So, PET is an imaging technique which can visualize metabolic functions.

A slice of a 3D PET image of a male patient with prostate cancer is shown in Figure 2.10. High signal intensities correspond to high levels of radiation and are marked red. Those regions are for example the tumour in the prostate as well as several metastatsis outside the prostate. Because cancer tissue has a fast metabolism, those regions have a high signal intensity, whereas other areas composed of healthy tissue have a low signal intensity. In Figure 2.11, the histogram of the PET image intensity from one patient is shown. It can be seen that healthy tissue mostly has a signal intensity of $1 < SUV < 4$, whereas cancer tissue occupies a larger range of $4 < SUV < 18$.

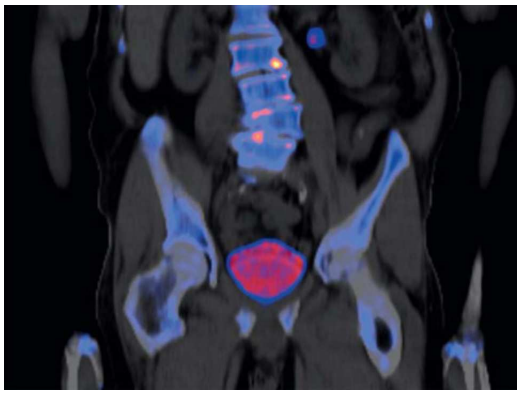

Figure 2.10.: 18F-NaF PET-CT showing precise delineation of bone metastases resulting from prostate cancer. Data courtesy of Northern California PET Imaging Center, Sacramento, CA, USA

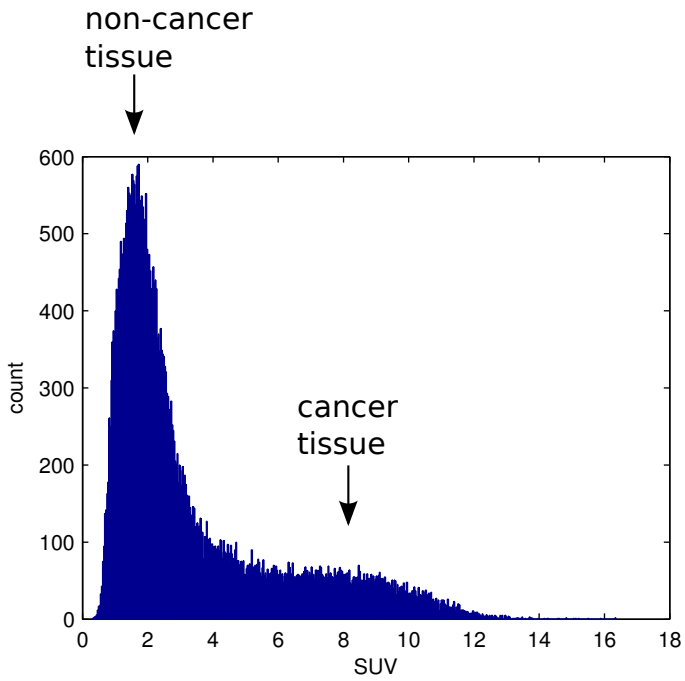

Figure 2.11.: Histogram of the SUV from one patient

The distribution shown in Figure 2.11 might experience a shift which is patient dependent.

The rate the tracer is absorbed in the tissue depends on the level of blood sugar. If the level of blood sugar is high, less tracer is absorbed by all of the tissue and therefore the distribution is shifted to the left. Another problem is that the rate of absorption depends on the time passed between injection of the tracer and measurement. Measurements are only comparable if the time between injection and measurement is the same.

Ground truth

The correct class labels for each voxel of a patient are the ground truth of the data set. For each patient there exists maps destinguishing between non-prostate tissue, healthy prostate tissue and cancer prostate tissue. As shown in Figure 2.12 for the patients 1-11, two maps composed of manual labels m_1 and m_2 are available for each patient. They have been created by cancer experts, analyzing the multispectral images. For patients 12-14, there exists one additional map m_3 for each patient derived from histological analysis of the prostate. Those patients underwent a surgery directly after measurement where the prostate has been extracted. The extracted prostate was cut into slices and the cell structure of extracted prostate was examined slice-wise under the microscope. This leads to a highly accurate label map which contains only few errors. In all label maps, the three different tissue types are encoded as:

$$m_k = \begin{cases} 0 & \text{non-prostate tissue} \\ 1 & \text{healthy prostate tissue} \\ 2 & \text{cancer prostate tissue.} \end{cases} \tag{2.3}$$

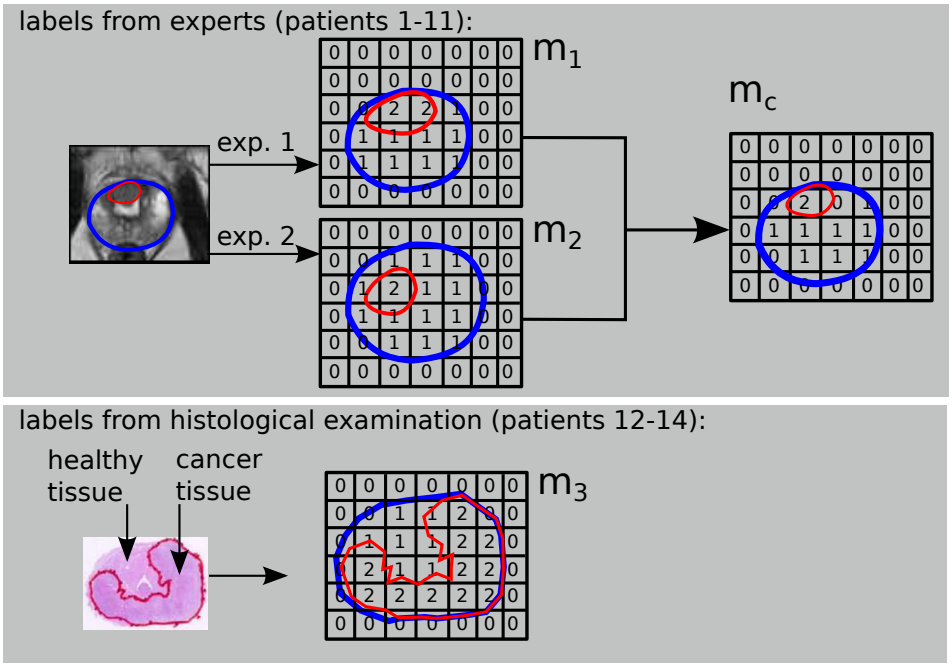

Figure 2.12.: Manual labels from two experts and histological analysis as ground truth

In general, the two maps from the cancer experts differ from each other. It is not ensured that either of those maps is correct, because manual delineation of cancer tissue is a very difficult

task which demands a lot of experience. Therefore, as shown in Figure 2.12, for the data sets 1-11 a combination m_c of the two expert maps m_1 and m_2 shall be used as ground truth

$$m_c = \begin{cases} 0 & \text{if } m_1 \neq m_2 \\ m_1 & \text{if } m_1 = m_2 \end{cases}. \tag{2.4}$$

This means, the expert labels are only trusted if both cancer experts agree. Because the histological analysis is more reliable, the maps m_3 derived from the histological analysis shall be used as ground truth for patients 12-14.

2.3. Tasks

The task of this PÜL is the implementation and training of classifiers for automatic prostate cancer segmentation. As described in Section 2.2, all patients contained in the medical data set are suffering from prostate cancer. So, the task is not to make a patient-wise decision (ill or healthy). The task is a voxel-wise classification of the prostate tissue into a cancer and a non-cancer class. So we want to know whether a specific voxel contains cancer tissue or not in order to get the exact location of the tumour.

Figure 2.13 gives an overview of the processing chain and tasks. All tasks which have already been done are drawn in light gray whereas all tasks still to be done are dark. At the end of the PÜL, each team has present in his own implementation of this chain together with his best trained classifier. Modifications or extensions to this processing chain may be introduced if they enhance performance or stability.

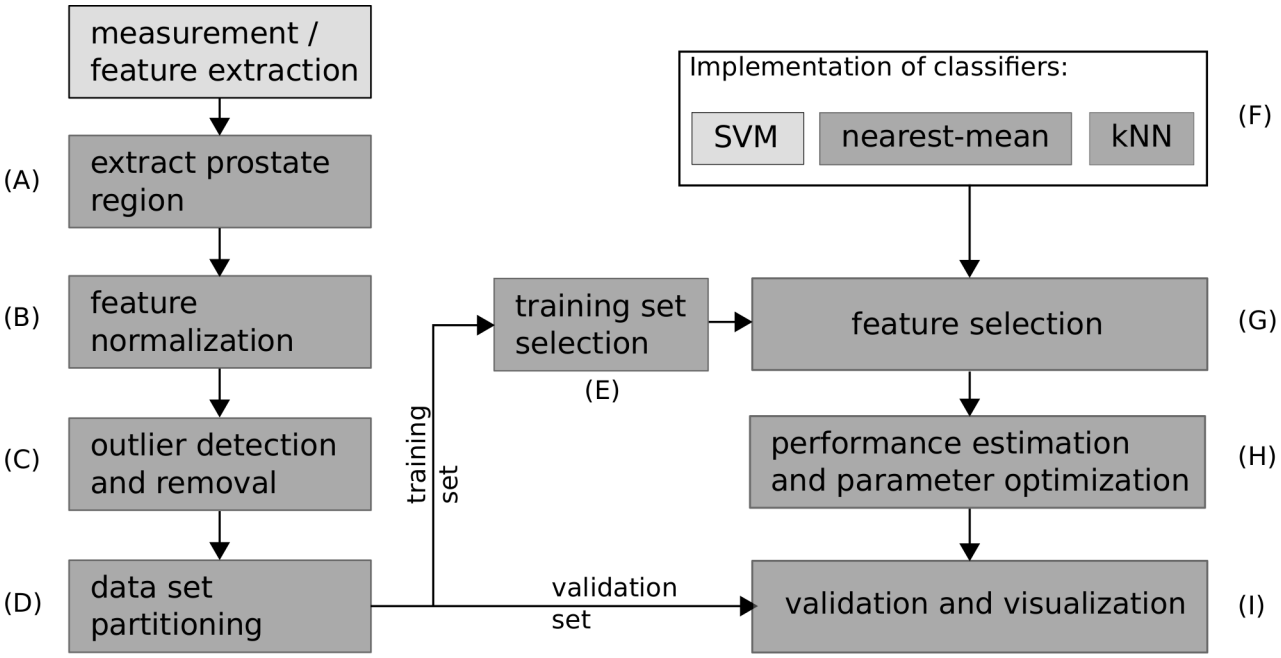

Figure 2.13.: The classification chain

2.3.1. Extract prostate region

Tissue in regions outside the prostate shall not be considered during classification. So feature vectors from regions with prostate tissue have first to be isolated from the whole data set. This can be done by using the maps described in Section 2.2.

A problem to be considered is that T2 weighted MR images are relative meassurements. So, T2 weighted images can not directly serve as features for cancer detection.

Task:

Implement functions to isolate feature vectors belonging to prostate tissue from the whole data set. Think of a way to transform a T2 weighted MR image such to make it a usefull feature for classification. This can ,for example, be achieved by scaling the dynamic range of the MR image to the range [0,1].

2.3.2. Feature normalization

To obtain good classification results, another preprocessing step is recommended after feature extraction. All features are obtained with different measurement techniques and describe different properties. This means that all features have different dynamic ranges. Depending on the classifier used, features with a large dynamic range may have a big influence on the decision boundary. This problem is discussed in details in [4], [11] and [12], together with an overview of the most common normalization.

Task:

Try to train the classifiers with different normalization techniques. Depending on the classifier in use, choose a method to scale the features which gives the best classification results. There are many different methods to scale features. However, there is no rule which one is suited best for specific applications. Commonly used normalization techniques are:

- Scale each dimension of the feature vector to zero mean and unit variance

$$\tilde{x}_k = \frac{x_k - \mu_k}{\sigma_k} \tag{2.5}$$

- Scale each dimension of the feature vector $\underline{x}$ to unit dynamic range [0,1]

$$\tilde{x}_k = \frac{x_k - \min(x_k)}{\max(x_k) - \min(x_k)} \tag{2.6}$$

- Scale the feature vector to unit vector length

$$\underline{\tilde{x}} = \frac{\underline{x}}{\|\underline{x}\|} \tag{2.7}$$

2.3.3. Outlier detection and removal

As mentioned in Section 2.2, the ground truth from the data set contains errors. Some classifiers are especially sensitive to wrongly labeled training data.

If, for example, a nearest-mean classifier is applied, the outliers in the training set will lead to wrong estimates of the class centres $\hat{\underline{\mu}}_k$ and therefore reduce the performance of the classifier. To avoid such problems, a so called outlier detection can be applied before training. The goal is to detect samples of the training and test set which are likely to have an erroneous label and to remove them from the database. Problems of such an approach are that during outlier detection, it is likely also to remove correctly labeled samples which lie near the decision boundary and therefore are important for the training of the classifier.

Task:

Implement the outlier detection method described in [13] and test it with artificially created two-dimensional distributions which contain outliers. Problems of the detector from [13] are that if many outliers occurr, effects known as masking and swamping will make the detection biased. This means that it is very likely that the algorithm will fail to detect the outliers.

Try to determine which percentage of outliers have to be expected in the given training set due to wrong manual labels. Is it likely that the outlier detection proposed in [13] can handle it? To enhance the robustness, you should search the literature for one improved method for outlier detection and compare the performance to the detector proposed in [13].

2.3.4. Data set partitioning

Training and testing of classifiers are based on the 14 multispectral data sets described in Section 2.2. In order to obtain comparable results, each team should split the data set equally. As shown in Figure 2.14, data sets 1 to 11 shall be used as training sets, whereas data sets 12 to 14 are test sets to assess the generalization performance of the trained classifiers. This is a reasonable choice, because an accurate ground truth to assess the performance of the classifier only exists for the data sets 12 to 14.

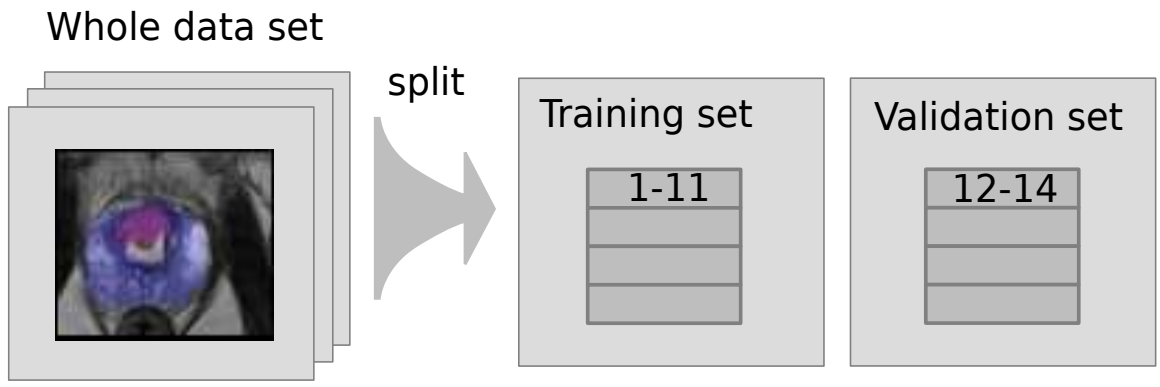

Figure 2.14.: Splitting of the data set

Task:

Design data structures to store the training and validation dataset. For training and validation of the classifiers, you need to implement methods to partition the given 14 datasets into a training set and a validation set by using these data structures.

2.3.5. Training set selection

As described in Section 2.2, each dataset for one patient contains approximately 35.000 voxels covering the prostate of the patient. Since the training set consists of 11 datasets, there are approximately 385.000 voxels for classifier training. Training a classifier with this vast amount of data is computationally complex and time consuming. To make a decision, a kNN classifier would need to calculate 385.000 distances per voxel, sort them by size and find the k smalles ones. If the size of the training set is not reduced, training the three classifiers may not be possible.

Task:

Algorithms for training set selection can be used to reduce the size of the training set while retaining the most critical training instances. The basic idea is that only training samples are used which lie near the decision boundary and therefore are important for the training of the classifier. The rest of the training samples can be discarded because they hold no additional information. Methods are described in [14] and [15]. You shall implement an algorithm to reduce the training set.

To show that the algorithm works properly, create an own test benchmark for a two-class classification problem using two-dimensional feature vectors. The algorithm for training set reduction works well if the decision boundary of a classifier trained with the reduced set will be similar to the decision boundary of a classifier trained with the whole training set.

2.3.6. Implementation of classifiers

Three different classifiers have to be compared for the task of cancer tissue segmentation. They are the nearest-mean classifier, the kNN classifier and the support vector machine (SVM). To get some insight in algorithms, you shall implement the first two classifiers by your self. The SVM has not to be implemented because the implementation is very challenging. Instead, the libSVM library shall be used which is available on the internet. Documentations of the libSVM are given in [16].

Nearest-mean classifier

The nearest-mean classifier describes each class by its class centre. An estimate of those class centres can be calculated by the sample average

$$\underline{\hat{\mu}}_k = \frac{1}{N_k} \sum_{n=1}^{N_k} \underline{x}_{k,n}, \tag{2.8}$$

where $\underline{\hat{\mu}}_k$ is the estimated class centre of class k and $\underline{x}_{k,n}$ are feature vectors of class k from the training set. For each new object with the feature vector $\underline{x}$, the distance

$$D(\underline{x}, \underline{\hat{\mu}}_k) = \sqrt{(\underline{x} - \underline{\hat{\mu}}_k)^T \hat{\mathbf{C}}_k^{-1} (\underline{x} - \underline{\hat{\mu}}_k)} \tag{2.9}$$

is calculated for each class centre. For the case that $\hat{\mathbf{C}}_k^{-1}$ is the estimated class covariance matrix, this is equal to the Mahalanobis distance, which is invariant to feature scaling. For the case $\hat{\mathbf{C}}_k^{-1} = \mathbf{I}$, it is equal to the Euclidean distance. The decision is based on finding the class centre which yields minimal $D(\underline{x}, \underline{\hat{\mu}}_k)$. Using the Mahalanobis distance, non-linear decision boundaries can be realized. If the Euclidean distance is used, only linear decision boundaries are possible.

kNN classifier

The kNN classifier is a so called lazy classifier. There is no need for computation during training phase because the classification algorithm operates directly on the training data. The algorithm is very simple and works as follows:

- For a given new feature vector $\underline{x}$, find the k nearest feature vectors out of the training set
- Decide for the class which is most oftenly represented among those k nearest neighbours

The parameter k controls the complexity of the decision boundary. If k is large, the boundary is smooth. If k is small, the decision boundary gets complex and the classifier tends to overfit. Classification with the kNN algorithm is very time consuming if the training set is large.

2.3.7. Feature selection

Due to the low number (only 5) of features for each voxel, the given classification task is computational not very demanding. Hence, reducing the number of features would not lead to a remarkable saving of computation time. However, acquiring all five images in Figure 2.3 is a time-consuming and costly process. Reducing the number of features means to reduce the measurement time. This makes the cancer diagnosis more cost-efficient and more comfortable for patients. Therefore, the most important features shall be identified and selected.

Task:

Implement a feature selection algorithm which chooses a subset of the most important features. Determine the best subset if only two of the five features shall be used for classification. You can either try all $\binom{5}{2}$ possibilities to get the best subset of features or you can implement other feature selection techniques as given in [4] or [17].

2.3.8. Performance estimation and parameter optimization

Some classifiers, for example the kNN algorithm, have parameters which effect the shape and complexity of the decision boundary and which should be chosen carefully. A wrong choice of those parameters may cause the classifier to overfit. This means the decission boundary may get too complex, yielding a very low training error rate but a high generalization error rate with the test data. It is always preferrable to choose the parameters of the classifier in order to obtain the lowest generalization error rate. Examples of such parameters is k of the kNN classifier, the regularization parameter C of the soft-margin SVM defining the penalty of wrongly classified samples or kernel parameters like γ if a polynomial kernel is used.

Task:

To train a classifier properly, an algorithm to tune the parameters of the classifier has to be implemented. Therefore, first make a literature search how to estimate the generalization error of a trained classifier without using the validation set. Keywords you can look after are for example “hold-out estimate”, “the jacknife method” or “k-fold crossvalidation” which are discussed in [4]. Afterwards, use a suited method of optimization to find the best parameters for each of the three classifiers yielding the lowest generalization error. A reference for methods of parameter optimization is [18].

2.3.9. Validation and visualization

The generalization performance of the trained classifiers shall be investigated with the validation data set. To compare the three classifiers, the result of the cancer segmentation shall be visualized.

Task:

The performance of the classifiers can be compared using Receiver Operating Characteristic (ROC) plots or by comparing error rates. The classification result shall be visualized appropriately. Regions of cancer tissue shall be marked in the T2 weighted image. For the case that a subset of two features is used for classification, the decision boundaries of all the three classifiers can be visualized in the two-dimensional feature space to show the differences.

3. Speaker recognition

3.1. Motivation

Today, there exist different techniques to automaically recognize the identity of a speaker from given voice recordings. Such automatic systems, as proposed in [19], [20] or [21], can reach a recognition performance of about 96 percent what makes them usable in many interesting applications.

One of the common applications for speaker recognition is speaker annotation. Speaker annotation means, that an algorithm is able to decide who has spoken at which time instant. Such an annotation comes in handy for recordings of interviews or meetings, as well as for intelligence applications. The annotation can for example be used for automated archivation of records in a database, enabling a person based search of the audio content.

Another scenario where speaker recognition became relevant in the last years is authentication. We normally use a lot of different systems where personal authentication is necessary, for example Email services, Computers, Online Banking and various websites like Ebay. Certainly, everyone can think about a situation where he could not remember a password or pin number in a crucial moment. It would be much easier if no one would have to remember complicated passphrases, but if the systems could identify the user on the basis of an acoustic fingerprint. Authentication over voice could, in the future, become a new option to improve the protection of our personal data.

Figure 3.1.: The worst 10,000 passwords [22]

The aim of this lab course is to implement a supervised closed-set, text independent speaker identification system. **Supervised** means that there are audio samples from each speaker to be identified, which can be used to train the classifier.

Closed-set means that only audio recordings from known speakers are used.

Text independent because the text which is spoken is unknown and not relevant for classification. So the user does not have to read an explicit text in order to be recognized, but can talk just about anything what comes to his mind.

Speaker identification means, that there is only a single speaker in each audio file. So, speaker identification boils down to a simple multi-class recognition problem without time dependence.

3.2. The TIMIT audio database

number of speakers:	overall: 630 training set: 462 validation set: 168
number of recordings per speaker:	10
average length of each recording:	3 seconds
audio type:	mono
sample frequency:	16 kHz
file format:	wav
language:	english

To train the speaker models and to test the recognition performance, audio samples from the TIMIT database will be used in this part of the PÜL. The audio samples in this database have been collected to test speech recognition systems but they are also useful for speaker recognition. The database contains samples from 630 speakers. All speakers are divided into two subset. A trainings set with 462 speakers and a validation set with 168 speakers.

For each speaker there are ten audio records of short english sentences with a average length of three seconds. All audio files have been acquired using a single channel 16 kHz audio recorder, in an environment with little noise and reverberation.

The path structure of the TIMIT database is depicted in Table 3.1. The file 'spkrinfo.txt' contains the metha-informations for all audio files, such as informations about the gender, the age, the spoken dialect and the identity of each speaker. The training set and the verification set are separated into the folders 'train' and 'test'. Both sets are divided into eight subsets 'dr1' to 'dr8'. Each subset contains multiple speakers with ten recorded audio files. All ten files of one speaker are located in one folder. The folder names follow the pattern shown in Figure 3.2. An example for a female speaker would be 'faks0'.

The text file prompts.txt contains all sentences spoken in the audio recordings. The file names of the wave files correspond to these sentences in prompts.txt.

Before the audio wave files can be processed, they have to be converted to standard wave file

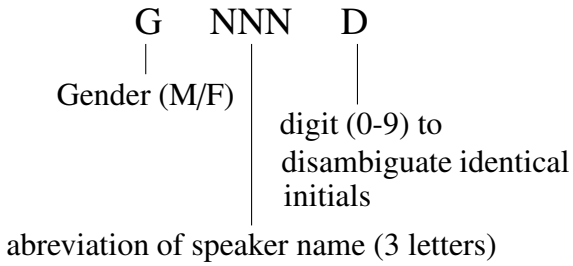

Figure 3.2.: folder name pattern

format. For this task, the provided tool 'sph2pipe' can be used, executed in the command line with the option '-f rif':

```
sph2pipe -f rif inputfile outputfile
```

```
timit
+     doc
|      +     prompts.txt
|      +     spkrinfo.txt
+     test
|      +          dr1
|      |           +          faks0
|      |           |           +       sa1.wav
|      |           |           +       sa2.wav
|      |           |           +       si943.wav
|      |           |           +          ...
|      |           +          fdac1
|      |           |           +       sa1.wav
|      |           |           +          ...
|      |          ...
|      +          dr8
|      |           +          ...
+     train
       +          dr1
       |           +          fcjf0
       |           |           +       sa1.wav
       |           |           +          ...
       |          ...
       +          dr8
                   +          ...
```

Table 3.1.: path structure of Timit database

During this part of the PÜL, each team is encouraged to acquire his own audio recordings in a special acoustic absorbent room in the basement of the ETI II (see Figure 3.3). Those recordings shall be used to test the recognition performance of the implemented software.

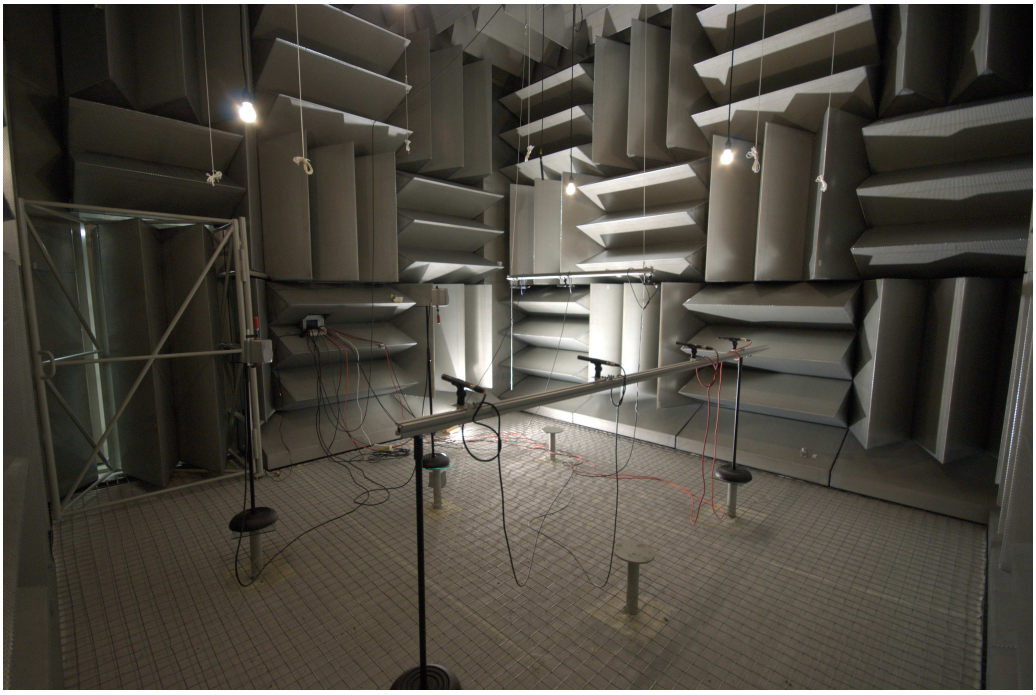

Figure 3.3.: Acoustic absorbent room

3.3. Tasks

3.3.1. The speaker identification pipeline

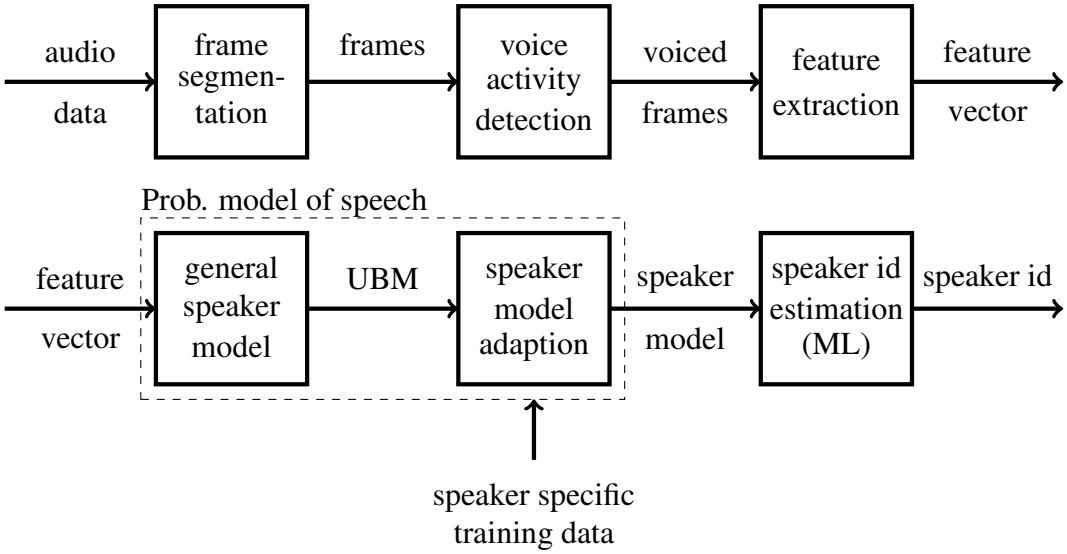

Figure 3.4.: Speaker identification pipeline

The block diagram of a basic speaker recognition pipeline consists is given in Figure 3.4. Its basic elements are:

- **frame segmentation:** In this preprocessing step the audio recordings are divided into smaller segments of length t_{frame} with feed t_{feed}. This segments are referred to as

frames.

- **voice activity detection:** Frames containing the actual speech signals have to be separated from frames of silence. Only the frames containing speech should be used to train the classification model. Otherwise the model would be trained on background noise.
- **feature extraction:** Features have to be extracted from the audio material which can be used to distinguish between multiple speakers. For this purpose MFCCs are widely used.
- **speaker model:** A general Gaussian Mixture Model (GMM) is used to capture the statistical distribution of the signal energy in the human voice spectrum. This GMM is trained on speech signals from different persons of different age and gender. It represents a general statistical model of human speech and, therefore, is called Universal Background Model (UBM).
- **speaker model adaption:** For each speaker to recognize, one Gaussian Mixture Model (GMM) is derived from this UBM. Therefore, the parameters of the UBM are adapted to cover the probabilistic distribution of the signal energy in the voice spectrum one specific person. For the adaption of the UBM only limited training data is needed, so each speaker only has to speak a short training sentence before the software can recognize him.
- **speaker estimation:** A maximum likelihood (ML) estimate is used to obtain the identity of the speaker.

3.3.2. Frame segmentation

The audio data has to be divided into small frames of length $t_{frame} = 20\ ms$ with feed $t_{feed} = 10\ ms$ between two neighbouring frames. Within these short time periods the signal can be assumed to be stationary. For the k-th frame with length t_{frame}, $L = \lfloor f_s \cdot t_{frame} \rfloor$ sample values x are taken from the audio data.

$$\underline{x}_k = \begin{bmatrix} x(k \cdot \lfloor f_s \cdot t_{feed} \rfloor) \\ x(k \cdot \lfloor f_s \cdot t_{feed} \rfloor + 1) \\ \vdots \\ x(k \cdot \lfloor f_s \cdot t_{feed} \rfloor + L - 1) \end{bmatrix} \tag{3.1}$$

with

$$k = 0, 1, \ldots, \frac{OverallNumberOfFrames - L}{\lfloor f_s \cdot t_{feed} \rfloor} \tag{3.2}$$

Figure 3.5 illustrates this segmentation of the audio material.

Tasks:

- Implement methods to read audio files from the TIMIT database.
- Create proper data structures to store the audio signals.
- Implement methods to sample frames from the audio signals.

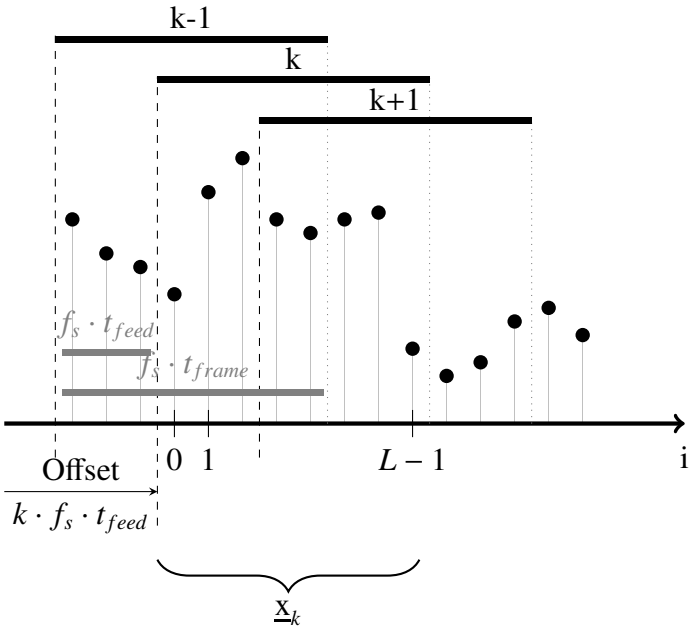

Figure 3.5.: Frame segmentation

3.3.3. Voice activity detection

In every speech there are small pauses when the speaker takes breath, waits for his or her dialogue partner or thinks about his next words. Extracting features from the unvoiced frames is not practical for speaker recognition, because they only contain background noise. To sort out these unwanted frames, a voice activity detection on basis of the signal and noise power has to be applied. So, only the voiced frames will be used later on to compute the features. Figure 3.6 shows a possible layout of an voice activity detector.

The individual steps of voice activity detection in detail:

- **step 1:** The individual signal and noise power $P(k)$ for the k-th frame is calculated by:

$$P(k) = \frac{1}{L} \sum_{i=0}^{L-1} x_k^2(i) \tag{3.3}$$

- **step 2:** It can be assumed that the first t_n seconds of every audio file contains no speech. The noise power P_N can, therefore be estimated from the first K frames:

$$K = \left\lfloor \frac{t_n}{t_{feed}} - 1 \right\rfloor \tag{3.4}$$

$$P_N = \frac{1}{K} \cdot \sum_{k=0}^{K-1} P(k) \tag{3.5}$$

- **step 3:** All frames with a signal and noise power greater than γ times the noise power are assumed to be voiced (hypothesis H_1) and all other frames as unvoiced (hypothesis H_0).

$$P(k) \underset{H_0}{\overset{H_1}{\gtrless}} \gamma \cdot P_N \tag{3.6}$$

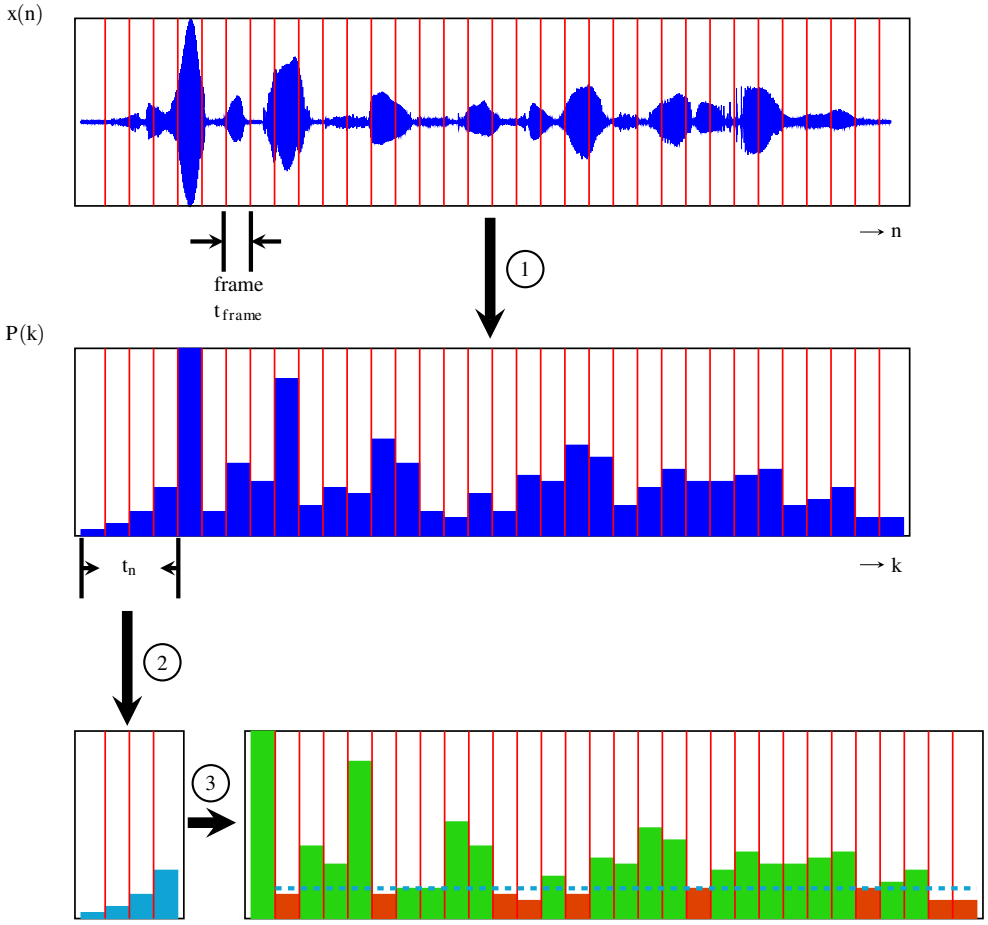

Figure 3.6.: Voice activity detection

Tasks:

- Implement a voice activity detector to separate voiced from unvoiced frames. The threshold has to be chosen carefully to allow a good detection rate.

3.3.4. Feature extraction

According to [23] the features used for speaker recognition should satisfy certain criteria:

- The features should be independent from what is spoken and should differ as much as possible between different speakers.
- They should be robust to background noise.
- They should occur frequently in normal speech.
- It should be hard to distort the features by disguising your voice
- and alteration of the voice by illness, tiredness, mood and age should not negatively affect the recognition performance.

Good features can be engineered by taking a closer look upon how the human voice is produced. Voice is formed by the human vocal tract which is depicted in figure 3.7

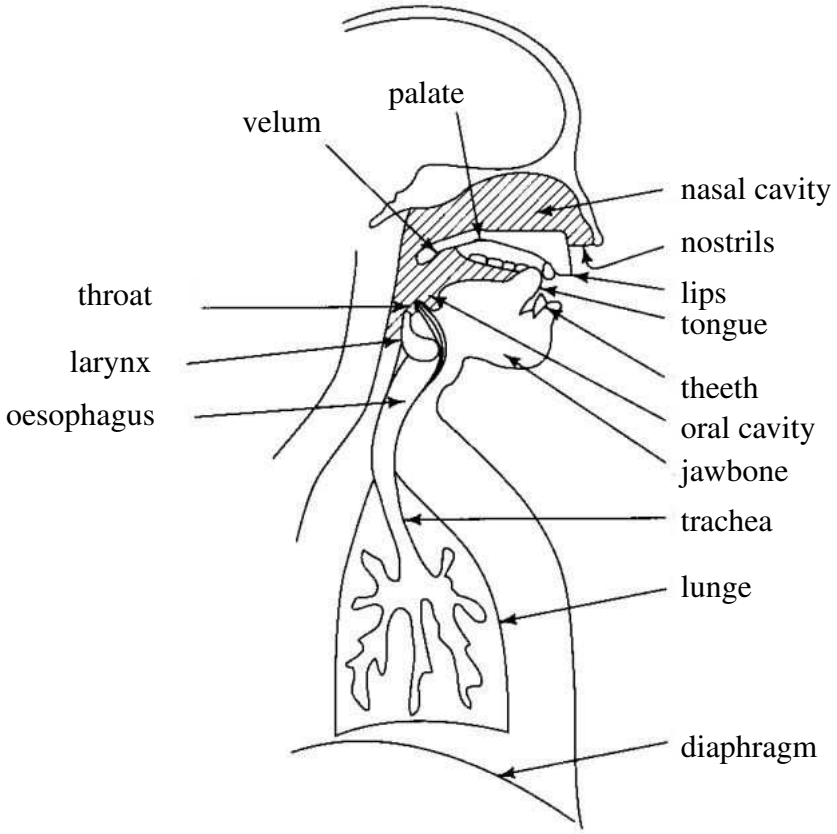

Figure 3.7.: human vocal tract

and consists of:

- the **larynx** with the vocal folds,
- the **throat or pharynx**,
- the **oral cavity** with tough, teeth and lips,
- and the **nasal cavity**.

The vocal folds regulate the air stream through the larynx. At vocal sounds the vocal cords in the vocal folds get excited. In the resonance chamber of oral and nasal cavity the voice spectrum gets altered. So excitation in the larynx and alteration in the oral and nasal cavity form our individual voice and and is mapped in the voice spectrum.

As the dimensions and properties of the vocal tract are different from person to person, everyone has a characteristic and unique voice. Our goal is to extract features from the spectrum which map these characteristic properties and allow us to differentiate between different persons.

In speaker recognition mainly two feature types based on the speech spectrum are applied, which satisfy these criteria. One are the Linear Predictive Cepstral Coefficients (LPCCs) and the others are the Mel Frequency Cepstral Coefficients (MFCCs), which will be used for this lab course.

Mel Frequency Cepstral Coefficients

According to the Weber-Fechner [24] law the human ear senses frequencies in a logarithmic way. This means lower frequencies can be discriminanted more accurately than lower frequencies. To utilizes this behaviour the spectrum is converted to a Mel spectrum with a logarithmic frequency axis defined by the Mel scale function [25].

Mel Scale

The Mel scale is a piecewise defined function:

$$f_{mel} = \begin{cases} f & \text{für} \quad f \leq 1\ kHz \\ 2595 \cdot \log_{10}(1 + \frac{f}{700}) & \text{für} \quad f > 1\ kHz \end{cases} \tag{3.7}$$

which is linear up to 1 kHz and logarithmic above.

Windowing

Due to the finite length of the frames, the will occur leakage effects after Fourier Transformation [26]. To minimize the effect, a proper windowing function has to be applied.

Mel Filter Bank

To convert the spectrum to Mel scale spectrum a Mel filter bank can be used. The Mel filter bank consists of multiple triangular shaped bandpass filters with there bandwidth defined by the Mel scale function. For the speaker recognition system a filter bank of $M = 22$ filters can be used.

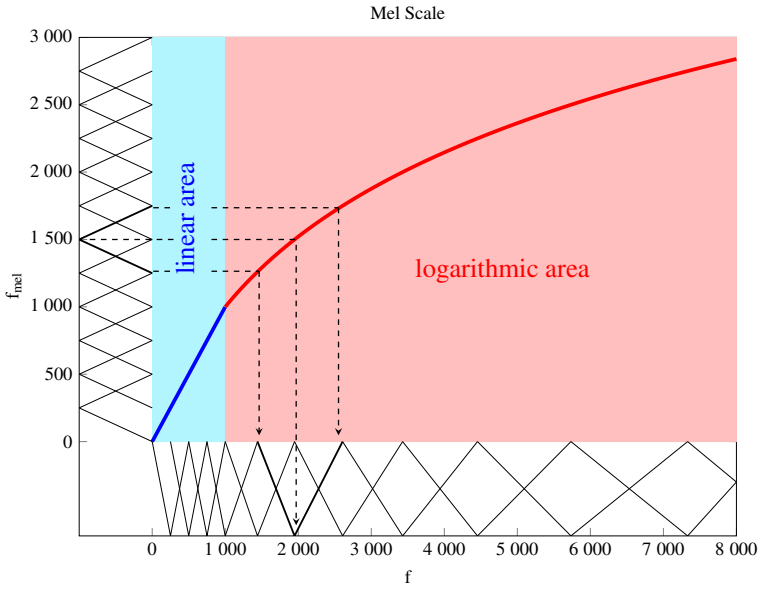

Figure 3.8.: Mel Filter Bank

To create the Mel filter bank you start with the wanted number of filters in the Mel frequency domain. All filters have the same bandwidth, neighbouring filters overlap by 50 percent and the filters are distributed over the desired frequency range. To get the filters in the normal frequency domain you use the inverse Mel Scale function and transform the filters into the normal frequency domain. Due to the logarithmic shape of the Mel Scale, filters in the lower frequency range have a lower band width than filters of higher frequencies. This leads to an increased weighting of the lower frequencies. Figure 3.8 shows this method.

Due to the overlapping of the filters, the features from the Mel filter bank are highly correlated. To decorrelate the features a Discrete Fourier Transform (DCT) [19] is appled. As input for the DCT the sums $Y(m)$ of the triangular filters are used. The DCT values x_n with n = 1,...,15 will then be our values for the MFCC feature vector $\underline{x}$. The DC-offset from $n = 0$ and the DCT values above $n = 15$ will not be used. Equation (3.8) shows the DCT function:

$$x_n = \sum_{m=1}^{M} [\log_{10} Y(m)] \cdot \cos\left[\frac{\pi n}{M}\left(m - \frac{1}{2}\right)\right] \tag{3.8}$$

Tasks:

- **Mel Scale:** Write two functions to represent the Mel Scale function and the corresponding inverse function.
- **Mel filter bank:** Use the Mel Scale functions for a new function which computes the Mel filter bank matrix. The filter matrix has $M = 22$ rows and $\lfloor f_s \cdot t_{frame} \rfloor$ columns. Each column describes one triangular filter of the filter bank.
- **DCT:** Implement a function to compute the MFCC feature vector of a frame from the Mel filter bank output by using the Discrete Fourier Transformation (DCT).

3.3.5. Probabilistic model of speech

For each speaker a naive Gaussian Mixture Model (GMM) can be used to from a statistical model of the characteristic distribution of the MFCCs. A GMM for speaker λ is described by its probability density function (pdf):

$$p(\underline{x}|\lambda) = \sum_{k=1}^{K} w_k p(\underline{x}|\underline{\mu}_k, \mathbf{C}_k) \tag{3.9}$$

with the Gaussian normal distribution for the general D-dimensional case:

$$p(\underline{x}|\underline{\mu}, \mathbf{C}) = \frac{1}{(2\pi)^{\frac{D}{2}} |\mathbf{C}|^{\frac{1}{2}}} e^{-\frac{1}{2}(\underline{x}-\underline{\mu})^T \mathbf{C}^{-1} (\underline{x}-\underline{\mu})} \tag{3.10}$$

$\underline{\mu}$ denotes the expectation and $\mathbf{C}$ the covariance matrix of $\underline{x}$. The weights have to satisfy the condition:

$$\sum_{k=1}^{K} w_k = 1 \tag{3.11}$$

Every speaker is described by on GMM $\lambda = \{\underline{w}_k, \underline{\mu}, \mathbf{C}_k\}$. The parameters $\underline{w}_k$, $\underline{\mu}$ and $\mathbf{C}_k$ have to be estimated with a training algorithm from training data of the specific speaker.

Universal Background Model

To reduce the amount of training data and the computational cost for training, a Universal Background Model (UBM) can be used which is pretrained on a large data set.

Good speaker models require a huge amount of training data such that the GMM parameters can be estimated as accurately as possible. But a speaker who wants to enrol into the system normally does not want to speaker for more than a few seconds.

To get good speaker model with little training data use a general speaker model or universal background model (UBM). The idea is that all speakers have certain similarities and that a model can be created which describes these common features. The individual speaker models can then be derived from this general speaker model and some speaker specific training data. This UBM is trained once with a very huge amount of training data and is then used as a parameter initialization for the training of the speaker models.

Speaker models

According to [20] the speaker model adaption from the UBM is done by a weighting of the UBM with speaker specific training data $\mathbf{X} = \{\underline{x}_1, \underline{x}_2, \ldots, \underline{x}_t, \ldots, \underline{x}_T\}$.

$$Pr(i|\underline{x}_t) = \frac{w_i p_i(\underline{x}_t|\underline{\mu}_i, \boldsymbol{C}_i)}{\sum_{k=1}^{K} w_k p_k(\underline{x}_t|\underline{\mu}_k, \boldsymbol{C}_k)} = \frac{w_i p_i(\underline{x}_t|\underline{\mu}_i, \boldsymbol{C}_i)}{p(\underline{x}_t|\lambda)} \tag{3.12}$$

With the mode $Pr(i|\underline{x}_t)$ and the training data $\mathbf{X}$, the expectation values $\underline{\mu}$ and the covariance matrix $\mathbf{C}$ of the new speaker model can be calculated.

The weighting of the UBM to the training data is calculated by:

$$n_i = \sum_{t=1}^{T} Pr(i|\underline{x}_t) \tag{3.13}$$

$$\alpha_i = \frac{n_i}{n_i + r} \tag{3.14}$$

where the relevance factor r can be chosen freely. A smaller relevance factor reduces the influence of the speaker specific training data and increases the weighting of the UBM.
The expectation value $\underline{\hat{\mu}}_i$ for the training data $\mathbf{X}$ is calculated by:

$$\underline{\hat{\mu}}_i(\boldsymbol{X}) = \frac{1}{n_i} \sum_{t=1}^{T} Pr(i|\underline{x}_t)\underline{x}_t \tag{3.15}$$

With the factor α_i the expectation factor for the new speaker model can be computed:

$$\underline{\mu}_i = \alpha_i \underline{\hat{\mu}}_i + (1 - \alpha_i)\underline{\mu}_{i,UBM} \tag{3.16}$$

The expectation value $\underline{\mu}_i$ can then be used to calculate, with the training data, the relation matrix

$$\hat{\boldsymbol{R}}_i = \frac{1}{n_i} \sum_{t=1}^{T} Pr(i|\underline{x}_t)\underline{x}_t \underline{x}_t^T \tag{3.17}$$

and with that the new covariance matrix

$$\boldsymbol{C}_i = \alpha_i \hat{\boldsymbol{R}}_i + (1 - \alpha_i)(\boldsymbol{C}_{i,UBM} + \underline{\mu}_{i,UBM}\underline{\mu}_{i,UBM}^T) - \underline{\mu}_i\underline{\mu}_i^T \tag{3.18}$$

The weights for the speaker GMM are calculated by:

$$\hat{w}_i = \left(\frac{\alpha_i n_i}{T} + (1 - \alpha_i) w_{i,UBM}\right)\gamma \tag{3.19}$$

where γ has to fulfill the condition:

$$\sum_{i=1}^{M} \hat{w}_i = 1 \tag{3.20}$$

We have now calculated the new weights $\hat{w}_i$, the expectation values $\underline{\mu}_i$ and the covariance matrix $\mathbf{C}_i$ for the new speaker GMM.

Tasks:

- Write a function to derive a speaker model from the given UBM using the formulas from above and the speaker training data from the TIMIT database.

3.3.6. Speaker identification

In the speaker identification problem the correct speaker ω from a set of known speakers $\Omega = \{\omega_1, \omega_2, \ldots, \omega_K\}$ shall be assigned to a given set of feature vectors $\mathbf{X} = \{\underline{x}_1, \underline{x}_2, \ldots, \underline{x}_t, \ldots, \underline{x}_T\}$ from an audio sample. Every speaker is represented by a GMM λ_k and every speaker is equally likely. The most likely speaker $\hat{\omega}$ is estimated, as in [27] described, by calculating the concatenated probability:

$$\log P(\boldsymbol{X}|\lambda_k) = \sum_{t=1}^{T} \log p(\underline{x}_t|\lambda_k) \tag{3.21}$$

and taking the speaker of the most likely model:

$$\hat{\omega} = \arg\max_{k\in\Omega}(\log P(\boldsymbol{X}|\lambda_k)) \tag{3.22}$$

for $p(\underline{x}_t|\lambda_k)$ see equation (3.9).

Tasks:

- Write a function to assign one speaker model of a set of known speaker models to a set of feature vectors which are extracted for the audio sample for which you want to estimate the speaker.

Bibliography

[1] Yang, B.: *Detection and pattern recognition: Lecture notes and video recordings*. University of Stuttgart, 2014

[2] Niemann, H.: *Klassifikation von Mustern*. 2nd. Springer, 2003

[3] Richard O. Duda, David G. S. Peter E. Hart H. Peter E. Hart: *Pattern Classification*. 2nd. John Wiley & Sons, 2000

[4] Andrew R. Webb, Keith D. C.: *Statistical Pattern Recognition*. 3rd edition. John Wiley & Sons, 2011

[5] Bishop, Christopher M.: *Pattern Recognition and Machine Learning*. Springer, 2006

[6] Y, Peng ; Y, Jiang: Quantitative analysis of multiparametric prostate MR images: differentiation between prostate cancer and normal tissue and correlation with Gleason score–a computer-aided diagnosis development study. In: *Radiology* 267(3) (2013), S. 787–96

[7] Moradi, M. ; Mousavi, P. ; Boag, A.H. ; Sauerbrei, E.E. ; Siemens, D.R. ; Abolmaesumi, P.: Augmenting Detection of Prostate Cancer in Transrectal Ultrasound Images Using SVM and RF Time Series. In: *Biomedical Engineering, IEEE Transactions on* 56 (2009), Sept, Nr. 9, S. 2214–2224. http://dx.doi.org/10.1109/TBME.2008.2009766. – DOI 10.1109/TBME.2008.2009766. – ISSN 0018–9294

[8] Tofts, Paul S.: Modeling tracer kinetics in dynamic Gd-DTPA MR imaging. In: *Journal of Magnetic Resonance Imaging* 7 (1997), Nr. 1, 91–101. http://dx.doi.org/10.1002/jmri.1880070113. – DOI 10.1002/jmri.1880070113. – ISSN 1522–2586

[9] Tofts, Paul S.: T1-weighted DCE Imaging Concepts: Modelling, Acquisition and analysis. In: *Magnetom Flash* 3 (2010), S. 30–39

[10] Suetens, Paul: *Fundamentals of Medical Imaging*. Cambridge University Press, 2002

[11] Tax, David M. ; Duin, Robert P.: *Feature Scaling in Support Vector Data Descriptions*. 2000

[12] Aksoy, Selim ; Haralick, Robert M.: Feature normalization and likelihood-based similarity measures for image retrieval. In: *Pattern Recognition Letters* 22 (2001), Nr. 5, 563 - 582. http://dx.doi.org/http://dx.doi.org/10.1016/S0167-8655(00)00112-4. – DOI http://dx.doi.org/10.1016/S0167–8655(00)00112–4. – ISSN 0167–8655. – Image/Video Indexing and Retrieval

[13] Rousseeuw, Peter J. ; Zomeren, Bert C. v.: Unmasking Multivariate Outliers and Leverage Points. In: *Journal of the American Statistical Association* 85 (1990), Nr. 411, pp. 633-639. http://www.jstor.org/stable/2289995. – ISSN 01621459

[14] Almeida, M. Barros d. ; Padua Braga, A. de ; Braga, J.P.: SVM-KM: speeding SVMs learning with a priori cluster selection and k-means. In: *Neural Networks, 2000. Proceedings. Sixth Brazilian Symposium on*, 2000. – ISSN 1522–4899, S. 162–167

[15] Hart, P.: The condensed nearest neighbor rule (Corresp.). In: *Information Theory, IEEE Transactions on* 14 (1968), May, Nr. 3, S. 515–516. http://dx.doi.org/10.1109/TIT.1968.1054155. – DOI 10.1109/TIT.1968.1054155. – ISSN 0018–9448

[16] Chih-Wei Hsu, Chih-Jen L. Chih-Chung Chang C. Chih-Chung Chang: LIBSVM: A Library for Support Vector Machines. In: *none* 1 (2013), S. 39

[17] Guyon: An Introduction to Variable and Feature Selection. In: *J. Mach. Learn. Res.* 3 (2003), März, S. 1157–1182. – ISSN 1532–4435

[18] Chih-Wei Hsu, Chih-Jen L. Chih-Chung Chang C. Chih-Chung Chang: A Practical Guide to Support Vector Classification. In: *none* 1 (2003), S. 1–16

[19] Tomi Kinnunen, Haizhou L.: An Overview of Text-Independent Speaker Recognition: from Features to Supervectors. (2009), July, 30. http://cs.joensuu.fi/pages/tkinnu/webpage/pdf/speaker_recognition_overview.pdf

[20] Douglas A. Reynolds, Robert B. D. Thomas F. Quatieri Q. Thomas F. Quatieri: Speaker Verification Using Adapted Gaussian Mixture Models. In: *Digital Singal Processing* 10 (2000). http://speech.csie.ntu.edu.tw/previous_version/Speaker%20Verification%20Using%

[21] Alfredo Maesa, Michele Scarpiniti Roberto C. Fabio Garzia G. Fabio Garzia: Text Independent Automatic Speaker Recognition System Using Mel-Frequency Cepstrum Coefficient and Gaussian Mixture Models. In: *Journal of Information Security* (2012), March

[22] https://xato.net/wp-content/xup/passwordscloud.png

[23] Kinnunen, H. L. T.: *An overview of text-independent speaker Recognition: form features to supervectors*. http://cs.joensuu.fi/pages/tkinnu/webpage/pdf/speaker_recognition_overview.pdf Version: 2009

[24] *Weber-Fechner law*. Wikipedia. https://en.wikipedia.org/wiki/Weber%E2%80%93Fechner_la

[25] Hoang Do, Alex A. Ivan Tashev T. Ivan Tashev: A new speaker identification algorithm for gaming scenarios. (2001)

[26] *spectal leakage sine*. online. https://upload.wikimedia.org/wikipedia/commons/thumb/7/7 Version: September 2015

[27] Douglas A. Reynolds, Richard C. R.: Robust Text-Independent Speaker Identification Using Gaussian MMixture Speaker Models. In: *IEEE Transactions on Speech and Audio Processing* 3 (1995), Jan, Nr. 1. www.cs.toronto.edu/~frank/csc401/readings/ReynoldsRose.pdf

[28] Dhawan, Atam P.: *Medical Image Analysis*. John Wiley & Sons, 2003

A. Medical imaging techniques

The informations given in the following sections are not relevant to solve the classification task. They are only given for matters of completeness.

A.1. Magnetic Resonance Imaging (MRI)

The theory of magnetic resonance imaging demands deep knowledge of physics. However, detailed knowledge of MRI is not necessary for this PÜL. Therefore, only a brief and simplified overview is given which is beneficial to understand the significance of the features.

To get an MR image of a patient, the patient is placed in a strong and very homogeneous static magnetic field $\underline{B} = B_0\underline{e}_z$ like it is shown in Figure A.1. In our case the magnetic flux of the static field points in the z-direction and has the amplitude B_0. Due to the strong static magnetic field, the tissue is magnetized. This is called net-magnetization $\underline{m}$. In the state of equilibrium, this net-magnetization points in z-direction

$$\underline{m} = m_0\underline{e}_z \tag{A.1}$$

and has the amplitude m_0.

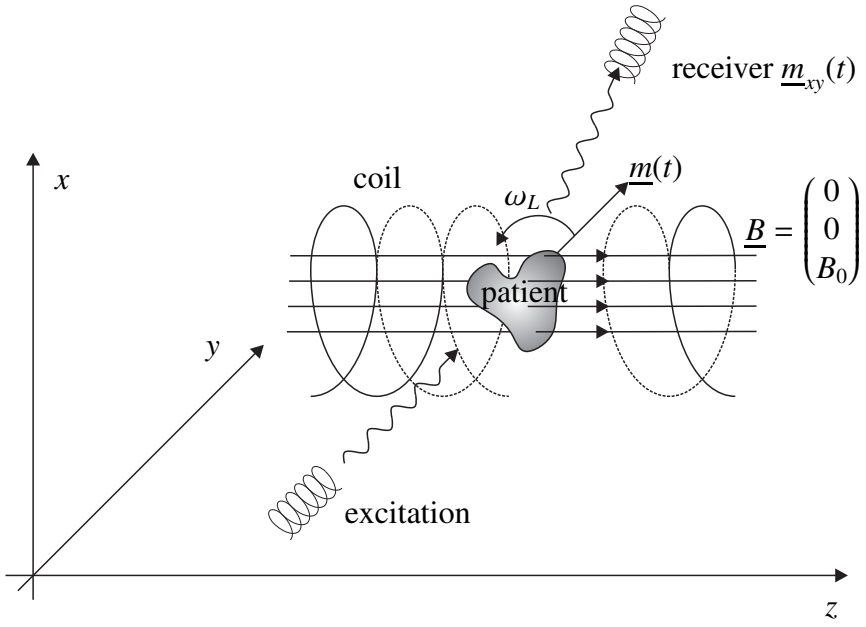

Figure A.1.: Principles of magnetic resonance imaging

If the tissue of the patient is excited with a second oscillating magnetic field whose field components are perpendicular to the static magnetic field

$$B_{RF}(t) = \begin{pmatrix} B_{RF}\sin(\omega_L t) \\ B_{RF}\cos(\omega_L t) \\ 0 \end{pmatrix}(u(t) - u(t - t_{ex})), \tag{A.2}$$

this state of equilibrium is distorted. B_{RF} is the amplitude and ω_L is the frequency of the RF pulse which is known as the Lamor-frequency. The pulse is turned on at the instant $t = 0$ and turned of at $t = t_{ex}$.

The excitation causes a time-varying net-magnetization

$$\underline{m}(t) = \begin{pmatrix} m_0\sin(\omega_{flip}t_{ex})\sin(\omega_L t)\exp\left(\frac{-t}{T2}\right) \\ m_0\sin(\omega_{flip}t_{ex})\cos(\omega_L t)\exp\left(\frac{-t}{T2}\right) \\ m_0\cos(\omega_{flip}t_{ex})\left(1-\exp\left(\frac{-t}{T1}\right)\right), \end{pmatrix} \tag{A.3}$$

The product $\phi = \omega_{flip}t_{ex}$ is called flip angle and defines how far the vector of the net-magnetization is rotated from the equilibrium during excitation. In practice pulses are used which flip the net-magnetization 90 or 180 degrees. The initial net-magnetization is m_0 and the relaxation constants T1 and T2 describe how fast the MR signal decays after excitation. T1 describes how fast the original z-component of $\underline{m}(t)$ is rebuild and T2 describes how fast the x- and y-component of $\underline{m}(t)$ decay after excitation. The resonance frequency of the tissue $\omega_L = \gamma B_0$ is called Lamor-frequency and depends on the strength of the static magnetic field. The gyromagnetic ratio γ is a constant which depends on the type of the tissue under investigation.

Using a receiver coil, the components

$$s(t) = m_0\sin(\omega_{flip}t_{ex})\sin(\omega_L t)\exp\left(\frac{-t}{T2}\right) + \mathrm{j}m_0\sin(\omega_{flip}t_{ex})\cos(\omega_L t)\exp\left(\frac{-t}{T2}\right) \tag{A.4}$$

which oszillate in the xy-plane can be measured after excitation and are referred to as the MR signal.

In Figure A.2, the net-magnetization is plotted over time. The excitation pulse is turned on at time $t = 0$. As a result, the net-magnetization in z-direction decreases and a rotating component in the xy-plane begins to grow. At time instant $t_{ex} = 38$, the excitation pulse is turned of and the relaxation begins. We can observe the exponential decay of the rotating magnetization components.

Imaging techniques

To get a three dimensional image of the MR signal intensity, the volume has to be split up into little volume elements, the so called voxels and the signal coming from each voxel has to be measured separately.

To get informations about the local signal strength methods like slice selective escitation, what means that only a thin layer of the patiens body is resonant and gives a signal $m(t)$, as well as frequency and phase encoding can be used. Therefore, a sequence of gradients to

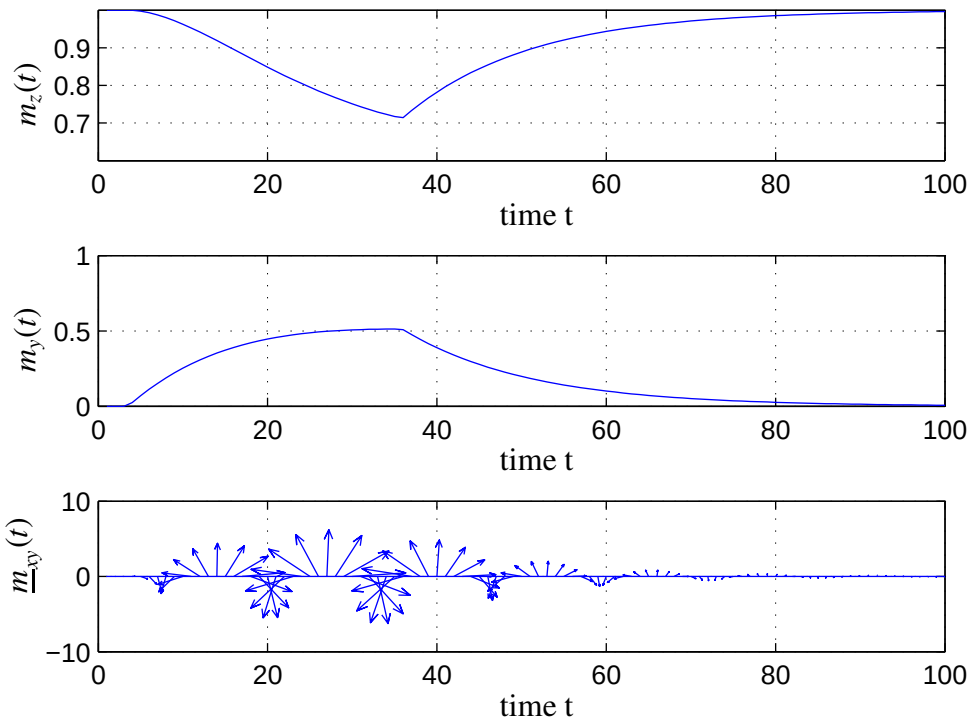

Figure A.2.: Net-magnetization versus time for an excitation with a RF pulse

the static magnetic field is applied during excitation and signal measurement. Due to those techniques, the measured signal is not only time dependant, but also has a spatial coding $s(t) \rightarrow s(x,y,z,t)$. Knowing the sequence of field gradients which have been used for measurement, an image containing the signal intensities of each voxel can be reconstructed from this signal. More detailed informations are containted in the books [28] and [10] which are available in the institutes library.

T2 weighted MR Images

In T2-weighted MR images, signal parts with long T2 relaxation time are weighted more than signal parts with short T2 relaxation time. In this way, the contrast of the image can be enhanced and different materials with different T2 can be distinguished.

The idea is to excite the tissue with an RF pulse and wait a specific echo time T_e before sampling the amplitude of the MR signal. What we measure for each voxel is the signal amplitude

$$|s(T_e)| = s(0)\mathrm{e}^{-\frac{T_e}{T2}}, \tag{A.5}$$

where $s(0)$ is the signal strength directly after the excitation pulse. Like depicted in Figure A.3, tissue (A) with long T2 has higher signal intensity than tissue (B) with short T2 which has already decayed further. Detailed informations about weighted imaging sequences are given in [10].

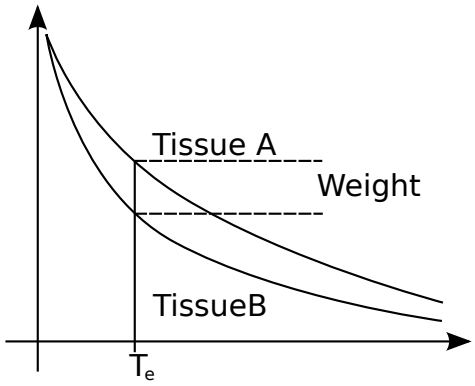

Figure A.3.: T2-weighting by measuring the signal intensity after the echo time T_e

Measurement of the apparent diffusion coefficient

To measure the ADC map, a special imaging sequence is used. As shown in Figure A.4, after excitation two similar gradients to the static magnetic field and one 180 degree rephasing pulse are used. The two gradients point into the direction in which the rate of diffusion shall be measured.

The idea is that the first gradient leads to a dephasing of the MR signal due to the spatial dependance of ω_L. After the first gradient has been applied, the nuclei either move through the tissue or stay at the same position. The 180 degree pulse and the second gradient are used to undo the dephasing caused by the first pulse. The dephasing is only reversed if the nuclei stay exactly at the same position (fixed) and each nuclei experiences the same effect of the gradient twice. The received MR signal then is strong. Dephasing can not be canceled if the nuclei move fast through the tissue (mobile). Each nuclei then will experience two different gradients and the spins will stay dephased and the received MR signal is weak.

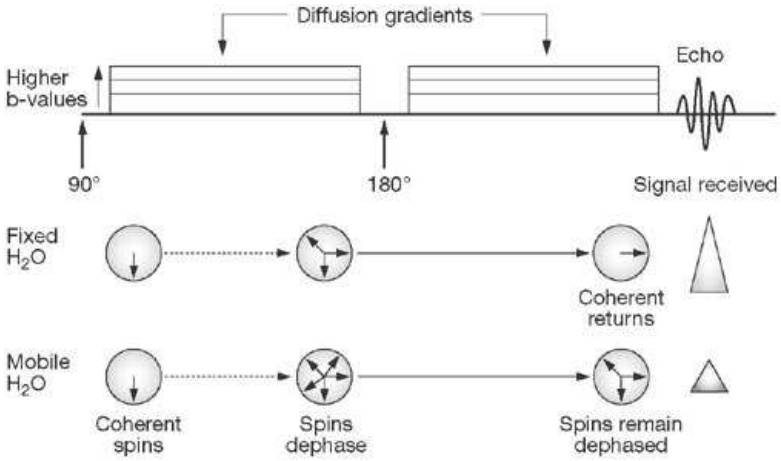

Figure A.4.: MRI sequence to measure an ADC map

The nuclei movement, therefore, leads to a faster T2-relaxation of the signal. The resulting

measured signal intensity can be written as

$$s'(t_m) = s(t_m) \exp\left(-\underbrace{(\gamma G \delta)^2 \Delta T}_{=b} D\right), \tag{A.6}$$

with the gradient amplitude G, the time δ the gradient is switched on, the time ΔT between the gradients, the diffusion parameter D and the instance of the measurement t_m. The signal $s(t_m)$ is the signal without dephasing due to the movement of the nuclei.

To calculate the diffusion coefficient D, the signal intensities $s'(t_m)$ are measured for different parameters b. Taking the logarithm of $s(t_m)$, the least squares method is used to calculate D from a sequence of measurements.

A.2. Positron emission tomography (PET)

The β^+-Emission

To measure the local radioactivity, PET relies on the so called positron emission (β^+-Emission). The injected tracer is a so called positron emitter. This means that protons are transformed into a neutron and a positron

$$p^+ \rightarrow n + e^+ \tag{A.7}$$

due to nuclear decay. This process is called positron emission and is shown in Figure A.5 on the left.

After each decay of a proton the newly formed positron travels a short distance (about a few millimetres) and slows down. When it has dissipated enough kinetic energy it can interact with a neighbouring electron which leads to the annihilation of both particles. This means that both particles are converted into energy. This energy is emitted in form of two photons with an energy of 511keV into opposite directions.

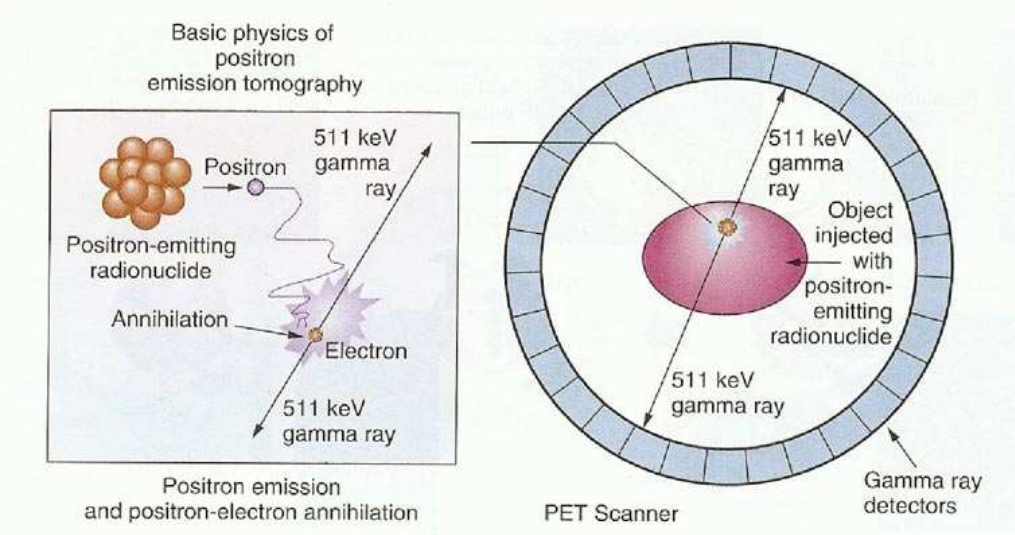

Figure A.5.: Principle of PET (Adopted from Molecular Imaging with Reporter Genes)

The detector of the PET is build to detect those photons and to estimate their origin. As shown in Figure A.5 on the right, detector cells are arranged in a ring around the patient. The basic layout of those detector cells is shown in Figure A.6. They consists of a scintillator which is sensible to photons with the specific energy of 511keV and which converts the photons into electrical pulses. This scintillator is mounted to the dynode of a photomultiplier tube (PMT) which amplifies the electric pulses. The signal at the output of the photomultiplier is used to detect and localize the β^+-emissions.

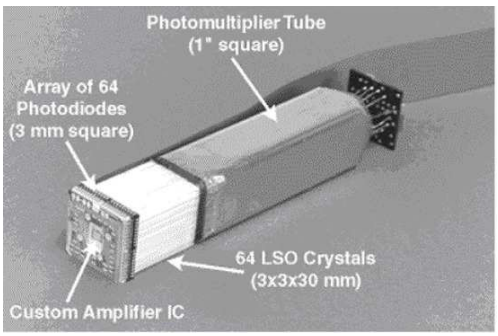

Figure A.6.: Example of a detector cell

Localization and Mapping

Because the two photons produced during electron/positron recombination are emitted into opposite directions, one knows that the origin of the β^+-emission has to lie on a straight line between two coinciding detections. The position of the β^+-emission on this line can be estimated if the time delay between two coinciding photon detections is known.

To get a map of the radio activity of the patients tissue, the patients body is divided into little cubic volume elements, the so called voxels. Single β^+-emission emitted from within the same voxel then are accumulated. The signal intensity of each voxel is equivalent to the absolute radioactivity $S_{PET}(t)$.

Standardized uptake value (SUV)

One problem is that we can not compare the raw PET measurements from different patients. The measured absolute local radio activity depends upon the amount of tracer and upon the activity of the tracer that has been injected. If the body mass of the patient is high the tracer has a lot of space to spread and therefore the measured local activity will be low in comparison with a PET measurement from a patient with less body mass. To ensure that measurements are comparable the so called standardized uptake value is defined which is a normalization of the measured activity

$$SUV(t) = \frac{S_{PET}(t)m}{a_i(t)}, \tag{A.8}$$

where $SUV(t)$ is the uptake value at the time instant of the measurement t, $S_{PET}(t)$ is the measured tissue radioactive concentration at time instant t, m is the body mass, $a_i(t)$ is the extrapolated total activity of the injected tracer at time instant t which varies due to nuclear

decay. If the tracer spreads equally distributed into the tissue, $SUV(t)$ will be one everywhere. At regions where the tracer is concentrated, $SUV(t) > 1$. At regions where the tracer concentration is lower than in case of an equal distribution, $SUV(t) < 1$. This shall ensure comparability of measurements from different patients.