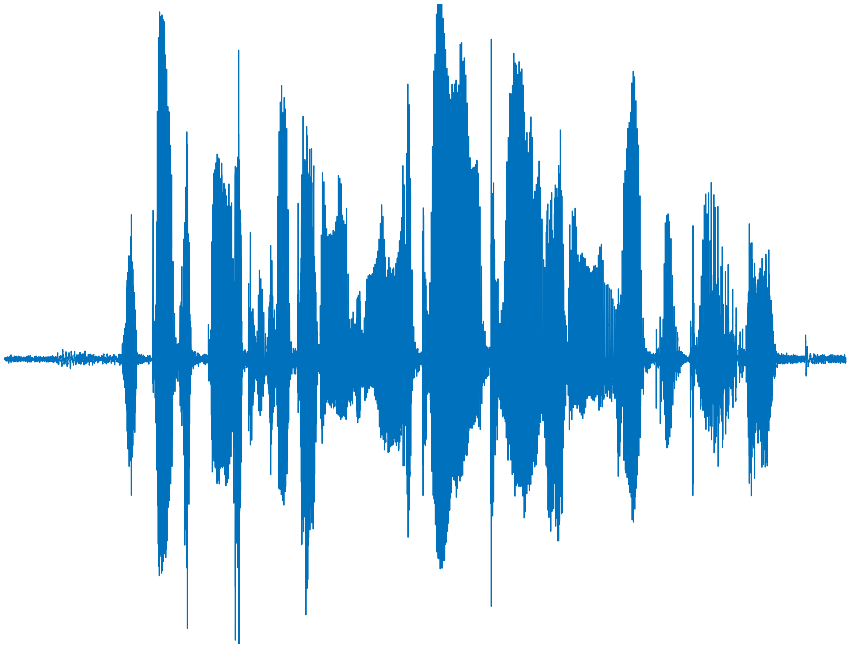
Algorithm based detection of pathological voices



Hampus Kalén

Master Thesis

Mathematical Statistics
Lund University
2018

Abstract

Automatic classification of pathological voices has the possibility of simplifying healthcare. In this project, classification of pathological voices from two different sources is attempted. Recordings of vowel pronunciations available in the Saarbrucken Voice Database (SVD) is used. Using features such as jitter, shimmer, Mel frequency Cepstral Coefficients (MFCC), and Harmonics-to-Noise ratio (HNR) in combination with Support Vector Machine (SVM) for classification, a correct classification rate of approximately 73 % is achieved. The other data set containing multiple recordings from the same patients reading short stories in different stages of vocal cord cancer is also examined. The sparsity of the data makes it difficult to make classifications, there are however indications that the MFCC's may be the most promising features for effective classification when more data is available.

Contents

Introduction

As with most types of cancer, the risk of reoccurring tumors is considerable. Today, patients need to revisit the hospital for follow up checks every few months. As cancer on the vocal cords influences the voice, often to a degree that is very well hearable, it is a reasonable to assume that a first assessment of the voice could be done by speech analysis, indicating that cancer is reappearing. The aim of this project is to implement an algorithm able to detect a change in the voice, possibly indicating reoccurring cancer.

A vision for the future would be that patients after surgery on the vocal cords could record their own voice, at home, at a regular interval, for instance, once per week. The recording would then be analyzed by a software that would detect at an early stage if there is a change in the voice. In this way, reoccurring cancer could be detected earlier, meaning the patient could get proper treatment earlier. This would both improve patient care and streamline the health care system.

1.1 Earlier research

Classification of pathological voices is a popular topic of ongoing research. The common approach is to extract a set of parameters, or features, from a recording and then use some sort of classifier to categorize voices as either healthy or pathological. The variation on which types of features and classifiers used is large, there is however some very commonly used. These include cycle-to-cycle perturbation measurements of the pitch and amplitude, jitter, and shimmer, respectively [1]–[6], harmonics-to-noise ratio (HNR) [3], [7]–[10], and mel frequency cepstral coefficients (MFCC) [2], [10], [11]. Besides these most common methods, researchers have used, for instance, linear prediction coefficients (LPC) and linear prediction cosine transform coefficients (LPCT) [9]. The common approach is to divide the voice signal into shorter frames of 10-100 ms and analyze each frame individually and also compare subsequent frames.

There is also a range of other types of parameters used from the temporal, spectral, cepstral, and correlation domain. Another approach, used in [12], [13], is to filter the voice signal to obtain the glottal flow and extract the relevant parameters. Discrete wavelet packet transform (DWPT) is another approach used in [11], [14].

The variation of methods used to classify the voice signals is even larger than the methods for parameter extraction. Some common classifiers used are support vector machines (SVM) with different kernels [3], [9], [10], Neural Networks [11], [14], k nearest neighbour (kNN) [14], and Gaussian mixture models (GMM) [8], [9].

The classification task of pathological voices have been completed rather successfully and many obtain a correct classification rate (CCR) above 90 % [9], [13], [15]. Due to the variation of the data used in different research studies it is problematic to compare CCR between different research groups. The difficulty in comparing different studies is discussed in [16], which finds a number of problems in comparing different studies to each other. In addition, it only handles comparisons for studies using the MEEI database, not the problematic nature of using different data sources.

What types of pathology differs between studies, there is work done detecting Parkinson's disease in [3], [4], some studies use voices with different types of issues, or the pathology is undefined [9], [14], [15]. There are also studies specifically focused on detecting laryngeal cancer [12], [17]. It seems that there is not much research in the area of distinguishing different diseases from each other, but research is more focused on the pathological-healthy classification task. There is no obvious difference in methods to determine different types of diseases either. Indicating that research done with respect to one disease or just the general pathological case, is also of interest for detection of other, specific diseases.

The data used in research around pathological voice detection usually consists of recordings of sustained vowel phonation, e.g. a person saying /ah/ for a few seconds. There is also recordings of regular speech, these are however, not as often used in research. There are two databases commonly used, the Massachusetts Eye and Ear Infirmary (MEEI) database, which is commercial. The Saarbrücken voice database (SVD) is a freely available database also frequently used in pathological voice research. In addition to these two, there is a number of smaller databases collected specifically by individual research groups.

1.2 Speech data

The data used for the project is audio recordings of patients reading a short story, approximately 1 minute long. The material includes one to nine recordings per patient ranging from pre-treatment recordings where there is some sort of problem with the vocal cords to post-treatment recordings when the vocal cords are fine. There is also recordings where the patient has some swelling from treatment or when there is a relapse of the illness.

Another source of speech data used here is a database called Saarbrücken Voice Database (SVD). It contains an extensive number of recordings collected from both healthy people, and people with various diseases effecting the vocal cords. An important difference in comparison to the recordings above is that SVD contains (among others) stationary recordings. Meaning that the person is pronouncing a vowel, instead of reading a short story. The different characteristics of the two data sets requires some different techniques in the analysis.

The idea is to extract quantitative characteristics, features, from the recordings that measures the characteristics that distinguishes between healthy and pathological voices. These features are then to be used to be be able to classify a new voice recording.

1.3 Objectives

There are multiple objectives this project aim to explore, all regarding detecting pathological voices. One angle is to analyze recordings from the same patients at different stages and see what differs between the recordings. A limitation with this approach is that the amount of data is very limited, making it more challenging to make persuasive classification. The available data for this aim only consists of non-stationary recordings, i.e., short stories.

Another angle is to explore if it is possible to extract features that distinguishes a pathological voice from a healthy voice, regardless of speaker. This objective comes with the difficulty that the voices of different people vary greatly, implying that the extracted features would need be independent of speaker specific traits, such as pitch among others. In exploring this objective the available data consists of both stationary and non-stationary data, i.e., both vowel pronunciations and short stories.

Speech data analysis

In this work, the analysis of the speech data can be divided into two stages. The first is to quantify a voice recording to a set of features describing relevant characteristics of the speech. The second step is to analyze and classify the data in the feature space.

2.1 Deriving features from voice data

2.1.1 Framing and pitch estimation

The voice signal is split up to 50 ms frames with 50 % overlap, 50 ms was chosen for two main reasons. Firstly, it was found to be a common duration in earlier studies of stationary voice. Secondly, for voices with a low pitch a longer time than, for instance 20 ms, which is another common frame length, was needed in order to obtain a sufficient amount of pitch periods. The pitch of each frame is then estimated using the YIN pitch estimator [18] with implementation provided in [19].

YIN is based on the autocorrelation method to find the fundamental frequency. The autocorrelation function (ACF) is computed by

$$r_t(\tau) = \sum_{j=t+1}^{t+W-\tau} x_j x_{j+\tau} \tag{2.1}$$

where x_t is the signal at time t, τ denotes a positive time lag, W the maximum lag for which the autocorrelation is computed. The simplest method to estimate the fundamental frequency is to choose the largest non zero lag peak in the ACF as the pitch period. Regrettably, this simple method, often results in errors. YIN uses a modified approach, beginning with a difference function which is somewhat related to the ACF and using a few more steps to improve accuracy. Assuming the signal is periodic with period T, then $x_t - x_{t+\tau} = 0$. Implying that the true period T should minimize the difference function

$$d_t(\tau) = \sum_{j=1}^{W} (x_j - x_{j+\tau})^2. \tag{2.2}$$

The difference function $d_t(\tau)$ will be zero for $\tau = 0$ and it will also be small for small τ, meaning that only minimizing $d_t(\tau)$ often results in pitch estimates that are too high frequency. To avoid this, one typically construct a new function that normalizes the difference function by its cumulative mean and removes the small values for small τ [18]

$$d'_t(\tau) = \begin{cases} 1, & \tau = 0 \\ \dfrac{d_t(\tau)}{\frac{1}{\tau}\sum_{j=1}^{\tau} d_t(\tau)}, & \tau > 0. \end{cases} \tag{2.3}$$

YIN makes use of a threshold of $1.1 \min(d'_t(\tau))$ to determine which candidates there is for the pitch period. The candidate corresponding to the smallest τ is then chosen as the pitch period. As a last step to get a more accurate estimate, a parabolic interpolation is done around the peak. This step is needed to be able to find a pitch period that is not a multiple of the sampling period. These steps are done both forward and backward on the signal and the best estimate is determined by which modified difference function contains the deepest dip.

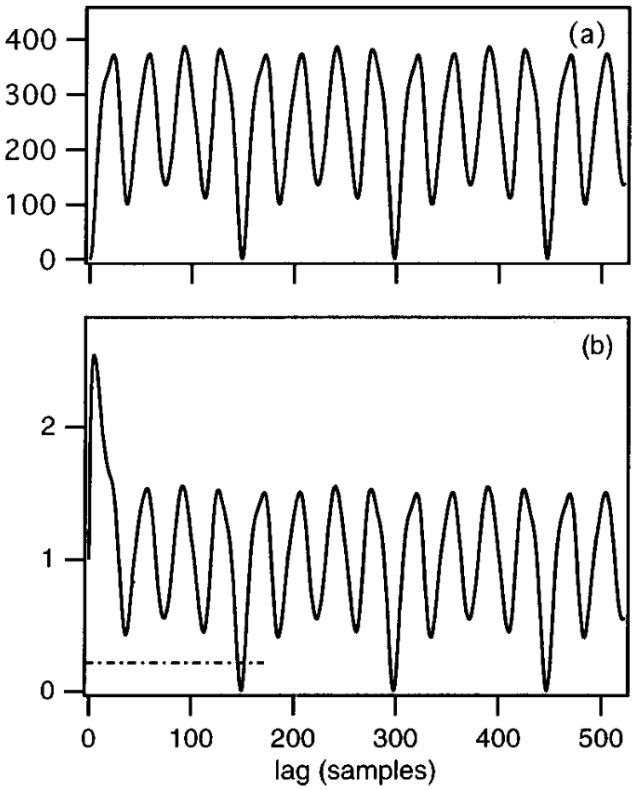


Figure 2.1: (a) The difference function $d_t(\tau)$ from equation 2.2 for a speech signal. (b) The modified difference function $d'_t(\tau)$ presented in equation 2.3 causing the original dip at zero lag to be one and increasing for small lags. The figure is from the original paper on YIN [18].

An issue with using YIN is that the estimate $\hat{f}$ of the pitch is often half or double the true pitch. To combat this problem a verification method using N overtones was developed. This method checks if half or double the estimate would make a better estimation. To check if half the estimate is the true frequency, the sum of the power of the half integer overtones is evaluated $P_{1/2} = \sum_{n=0}^{N} A_{(n-1/2)\cdot\hat{f}}$. This is then compared to $P_1 = \sum_{n=0}^{N} A_{n\hat{f}}$, where A_f denotes the amplitude at frequency f. If more than $P_{1/2} > 0.1P_1$ the pitch estimate is changed from $\hat{f}$ to $0.5\hat{f}$. To check if the original estimate should be doubled a similar method is used, the power spectrum is evaluated at even overtones, including the pitch, as well as odd overtones. $P_{odd} = \sum_{n=0}^{N} A_{(2n+1)\cdot\hat{f}}$, $P_{even} = \sum_{n=0}^{N} A_{(2n)\cdot\hat{f}}$. If $P_{odd} > 50P_{even}$, the even overtones were deemed to be too small to be overtones and the pitch estimate is instead doubled, from $\hat{f}$ to $2\hat{f}$.

The values 50 and 0.5 are results of experimental work together with qualitative assessment of the pitch to find suitable thresholds for when the pitch is not estimated correctly. The number of overtones to use in the verification was found to give smallest pitch variation when $N = 1$. There is however one problem with using $N = 1$, and that is when the power in the real pitch is very small in comparison to the first overtone, as can be seen in figure 2.2. Since the fundamental frequency is not visible in itself, but rather in the periodicity of the overtones for higher frequencies, it would make sense to use more overtones in the verification for this case. However, experience show that using more overtones increase the standard deviation of the pitch. The problem with this "missing fundamental" therefore remains. The construction of special tests using the amplitude of the overtones with the goal of detecting this abnormality was attempted, with little success, as the tests resulted in a large number of false positives.

To verify that this extra verification layer improves the pitch estimation, 100 stationary voice recordings were analyzed. As the recordings are stationary, we expect that the pitch will not change drastically between frames. Estimating the pitch for each frame and computing the standard deviation provides a measurement for how much the pitch varies. The mean of the standard deviation for the 100 recordings is presented in table 2.1.

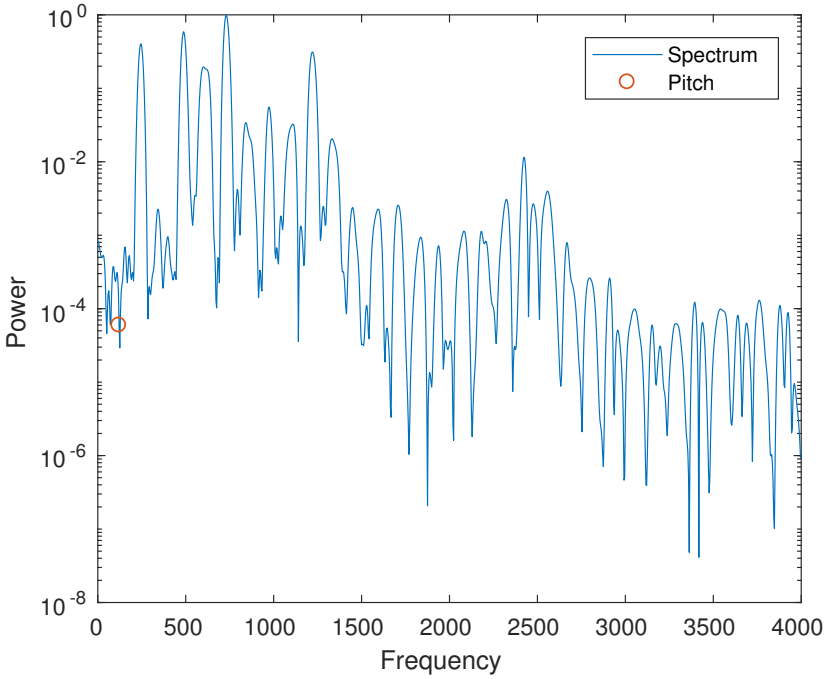


Figure 2.2: Spectrum of a 50 ms frame of a stationary voice recording. At first glance it seems that the pitch is wrongly estimated since there is virtually no power in the pitch, suggesting the large peak at approximately 250 Hz is the fundamental frequency. However, looking at the overtones above 800 Hz, it is evident that the pitch estimate is in fact correct. If not, the overtones would not appear as closely to each other as they do.

Table 2.1: Mean standard deviation of pitch within recording. Yin in combination with verification using the power in the overtones proves to give a more stable pitch estimate.

Method	Standard deviation
Only YIN	11.17
With verification method	4.56

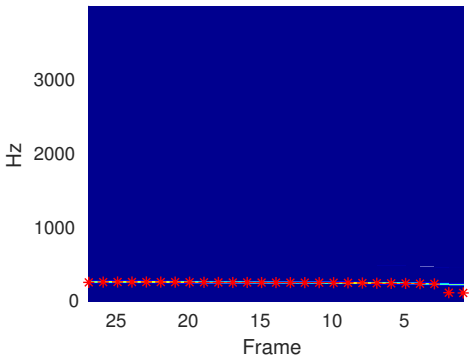
2.1.2 Extracting stationary voiced parts of the recordings

For general speech recordings, an important preprocessing step is needed. Since the vocal cords only are active when pronouncing a vowel, these frames of the recordings need to be extracted. The method used is to evaluate the minimum amplitude of the modified difference function of equation (2.3). By setting a threshold of the minimum amplitude required for a frame to be labeled as voiced, it is possible to tune how restrictive one wish to be, in labeling frames as voiced. Due to the relatively large number of frames in each recording, one may set the threshold to a small value, meaning that more voiced frames will be considered unvoiced but no unvoiced frames are labeled voiced. An appropriate threshold was qualitatively assessed to be 0.1. In figure 2.4, we see that this results in some candidate frames being discarded. However, due to the large number of frames, this is not an issue.

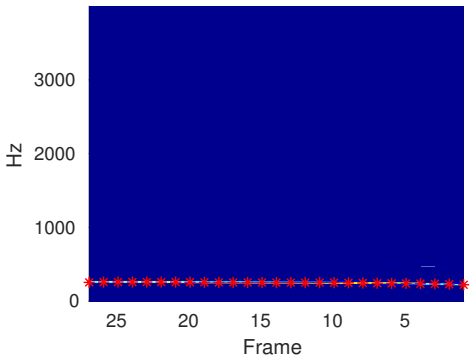
2.1.3 Jitter and Shimmer

A commonly used measurement used to separate pathological from healthy voices is jitter, which measures variations in the period between cycles. There are four different jitter values. First, there is jitta, a measurement of the average period variation in time, defined as

$$jitta = \frac{1}{N-1}\sum_{i=1}^{N-1} |\Delta T|, \tag{2.4}$$



(a) Spectrogram of a recording. Pitch estimation using only YIN shown with red stars. The estimation for the last two frames is incorrect.



(b) When using YIN in combination with the presented verification method the estimation for the last two frames is changed to be correct.

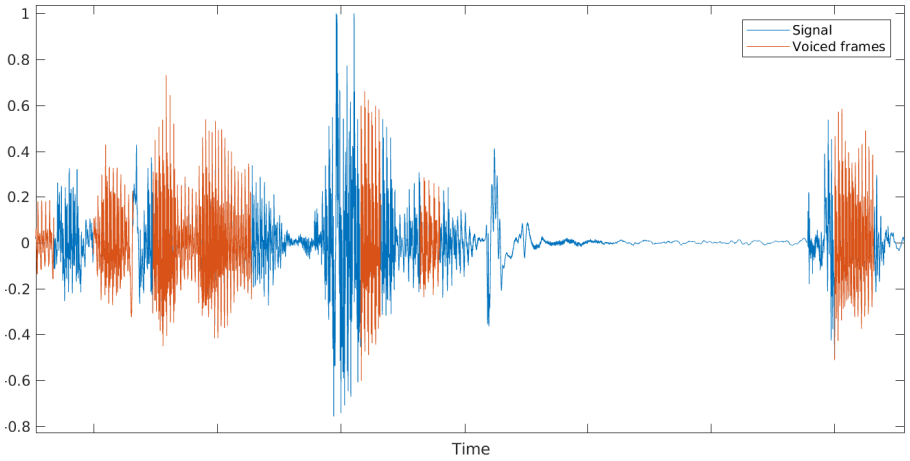


Figure 2.4: Part of an approximately one minute long voice signal. The voiced frames are displayed in orange.

where $|\Delta T| = |T_i - T_{i+1}|$ denotes the difference between two consecutive periods.

Secondly, there is jitt that is given as a percentage of jitta divided by the average period $\overline{T}$, defined as,

$$jitt = 100 \cdot \frac{jitta}{\overline{T}}. \tag{2.5}$$

Finally, Rap and ppq5 measures the difference between a period and the average of the adjacent periods, i.e.,

$$rap = 100 \cdot \frac{\frac{1}{N-1}\sum_{i=1}^{N-1} |T_i - \frac{1}{3}\sum_{n=i-1}^{i+1} T_n|}{\overline{T}} \tag{2.6}$$

$$ppq5 = 100 \cdot \frac{\frac{1}{N-1}\sum_{i=1}^{N-1} |T_i - \frac{1}{5}\sum_{n=i-2}^{i+2} T_n|}{\overline{T}}. \tag{2.7}$$

Shimmer is similar to jitter in the sense that in measures variations between periods, however, shimmer measures variation in amplitude instead. There is four different shimmer values: Shim, apq3, apq5, and ShdB. Shim, apq3, and apq5 are defined in the same way as jitt, rap, and ppq5, respectively, with the exception that it uses the amplitude A_i instead of the period T_i. ShDb however, is defined differently, as

$$ShdB = \frac{1}{N-1}\sum_{i=1}^{N-1} |20\log(A_{i+1}/A_i)|. \tag{2.8}$$

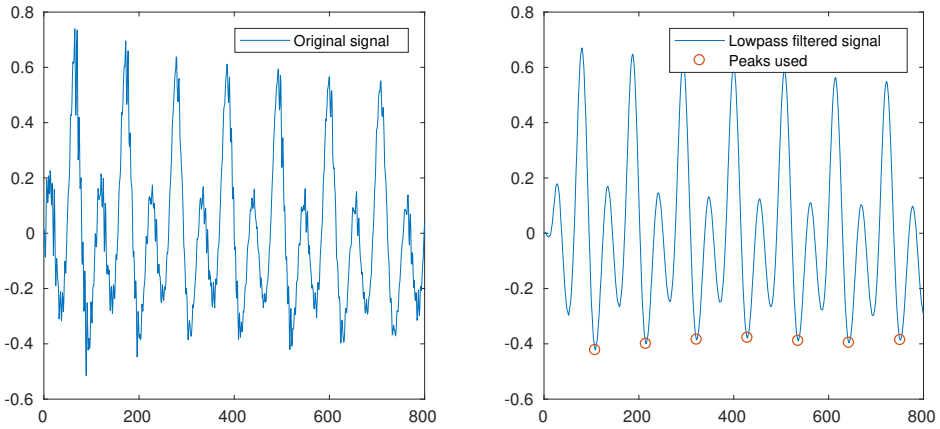


Figure 2.5: Using a lowpass filter on a noisy signal (left) enables finding the peaks (right) necessary to compute jitter and shimmer.

The problem of computing jitter and shimmer is essentially that of finding the peak of each glottal period. As the pitch has been determined earlier the period of the data is known. As a first step, to deal with noisy data, the signal is lowpass filtered, the filter is constructed to not alter the fundemental frequency but to remove high frequency noise in the signal. The peaks are then found by taking the max of the first period, the first peak, and then searching for the max value in a window the size of half a period centered around the next expected peak. This is done through the length of the data sequence for both positive and negative peaks (dips). Jitter and shimmer are then computed from the peaks with the smallest variation in periodicity. An example of this peak detection is shown in figure 2.5, where the need for the lowpass filtering is also displayed. For the computation of jitter and shimmer, the pitch needs to be estimated accurately, putting emphasis on the importance of pitch estimation.

2.1.4 Feature extraction in frequency domain

An important aspect of extracting information from the voice signal is by analyzing it in the frequency domain. Since the human voice exhibits an harmonic structure, finding the harmonic peaks is critical. This is done by

1. Finding the peak f_1 which corresponds to the fundamental frequency. This is done by picking the peak that are closest to the pitch estimate $\hat{f}$.
2. Find the subsequent peaks that is closest to the frequency $f_n = f_{n-1} + f_1$, up to a desired number of harmonics.

In figure 2.6, the spectrum of a 50 ms voice signal is shown, computed using a hamming windowed periodogram to suppress side lobes. Every peak in the spectrum, including noise peaks are displayed with a yellow circle. In addition, the fundamental frequency and the overtones are displayed in blue. The correct detection of the harmonics is of course greatly dependent on the pitch estimate to be valid, highlighting the importance of pitch estimation component.

The harmonics to noise ratio (HNR) is then computed by summing the power in the harmonic peaks P_h and dividing them by the power in the noise outside the peaks P_n, such that,

$$\text{HNR} = 10 \log \frac{\sum P_h}{\sum P_n}. \tag{2.9}$$

2.1.5 Mel frequency cepstrum coefficients

A common feature set used for categorizing voices both in speech recognition and detection of pathological voices are the mel frequency cepstral coefficients (MFCC) [2][20].

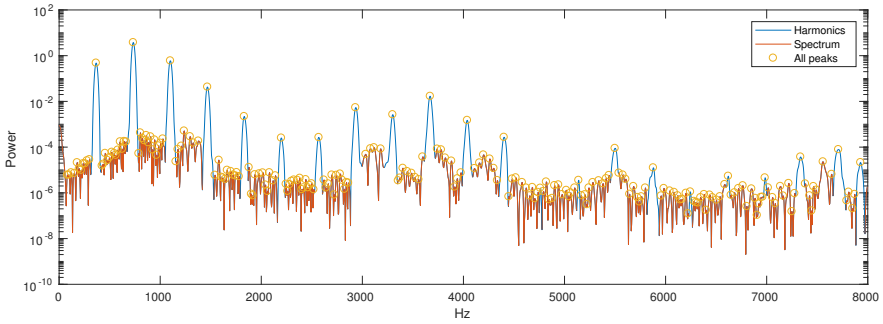


Figure 2.6: Spectrum of a 50 ms voice frame, all the peaks of the spectrum are found. The peaks corresponding to the first 20 harmonics is identified.

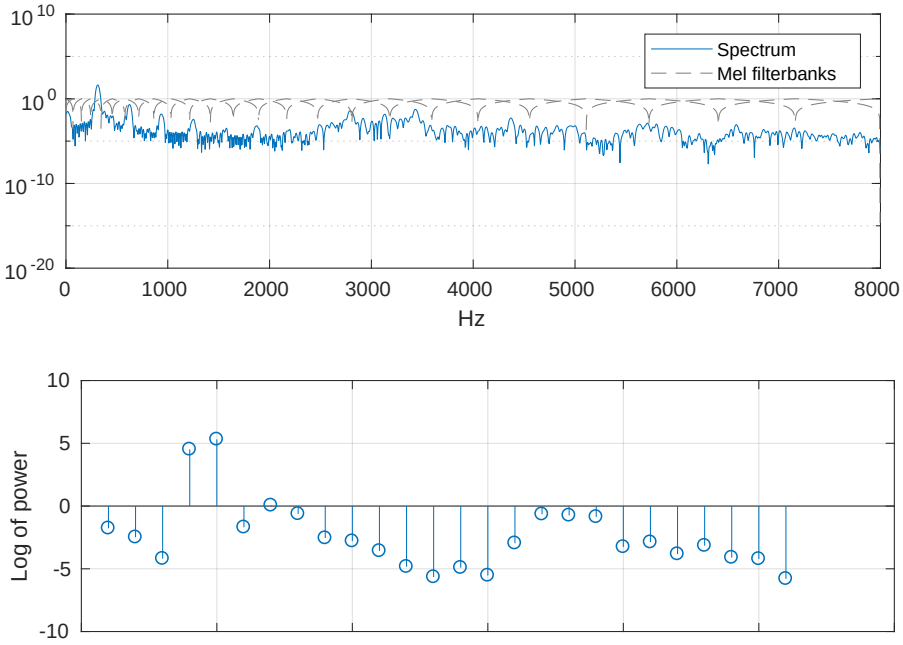


Figure 2.7: The upper figure shows the triangular filter banks that are equidistant in Mel scale. The logarithmic spacing of the frequency models the human hearing better than linear scale. The power in each of these filter banks is summed, which is shown in the lower figure. This data is then transformed using the discrete cosine transform to obtain the MFCCs.

Mel frequency is a logarithmic scaling of the frequencies, defined as,

$$m = 2595 \log\left(1 + \frac{f}{700}\right). \tag{2.10}$$

This scaling is done to better represent how well the human ear hears different frequencies. To obtain the MFCCs the FFT of the segment is summed over 26 equally Mel frequency spaced triangular filter banks with 50 % overlap. An illustration of the filter banks is shown in figure 2.7. The summed power over the filter banks is then cosine transformed to give the coefficients, i.e.,

$$C_k = \sqrt{2/N} \sum_{n=1}^{N} \log(P_n) \cos(\frac{\pi k(n - \frac{1}{2})}{N}). \tag{2.11}$$

2.2 Unused features

A couple more characteristics have been evaluated but have not been used as features. The reasons for discarding them varies, but some of the common ones are.

- Poor robustness, outliers very common.
- Difficult to quantify in a way compatible with the rest of the features.
- No indication that they may separate the two classes.
- Lack of earlier research using the feature.

The last item on the list raises the question, should a feature not be used simply because it has not been used by others? While the answers is no, the author chose to primarily focus on what has earlier been proven to work, before venturing into unexplored territory. With the exception if there are early indications that something new seems to work well.

A characteristic that was considered but not used is inharmonicity. Harmonic peaks means that the overtones are located roughly at $f_n = nf_1$, where f_1 is the pitch and n is an integer. However, due to the stiffness of the vocal cords, causing the frequency to shift somewhat from being purely harmonic it was briefly examined if a perturbation on the harmonic structure could be an indicator of pathological voice. The difference $\Delta_n = |f_n - nf_1|$ was evaluated. The perturbation on the harmonic structure was not used for a couple of reasons, the most important being that due to the limited resolution, the perturbation was almost always at most one frequency grid point. It may be worth examining further, for example by refining the grid around the harmonic peaks. Another reason why it was not examined as closely was simply that there were no indications that there is any difference between pathological voices and healthy voices.

Another feature considered was the width of the peaks. This was primarily discarded due to poor robustness and also the limited frequency resolution.

The amount of power in the signal for high and low frequencies was also examined by using complementary high- and lowpass filters. The issue with this was the vast variation between individuals. Another issue with such filtering is that it captures the power of both the harmonic and the noise part. An idea for developing this approach further is to first use a comb filter to remove the harmonic part and then analyze the remaining noise using a low- and highpass filter, or even filterbanks. With this method one could examine if the noise characteristics at different frequency bands could separate healthy and pathological voices.

2.3 Analyzing the feature space data

After the computations of, hopefully, relevant features presented in the previous section, the challenge of presenting them in a form suitable for further analysis and classification remains. This implies that all of the available voice data needs to be broken down into data points $\mathbf{x}_n$ with a corresponding label l_n indicating healthy/pathological. As we here compute one value for every feature, for each frame, a straightforward way would be to construct one data point for each frame, and assigning it the label of the entire recording. This would entail two effects, the first being the extensive amount of data points containing very similar information. The second implication is that the connection between frames from the same recording is lost, with the drawback that measuring longer variations in time would be impossible. One could circumvent this by categorizing frames belonging to one recording together, adding another layer of complexity. In conformity with the basic goal of the project, namely, to categorize a voice recording, and in the spirit of keeping things simple, it is here chosen that one recording should correspond to one data point.

Even after this choice of design have been made a number of challenges on how this is to be implemented remains. All computations presented in the previous section are done on each frame, the key part is to make use of the information of each frame in a manner to quantitatively represent the complete recording. In figure 2.8, the value of the feature ShdB for each voiced frame over the two recordings is shown. A few things may be noticed, the first being that there is no obvious difference in the distribution between the recordings. Furthermore, there is also big variation of a single feature over the course of a recording, with some outliers. The values of each feature is obviously of relevance, but the variation may also be of interest. To quantify this for the entire recording, the median over all frames was computed, as well as the standard deviation and the median of the frame-by-frame difference. This provides three different types of information: the absolute value, the variation over the entire recording, and the variation on a short time scale. The median is used as opposed to the mean due to the median being more robust with respect to outliers.

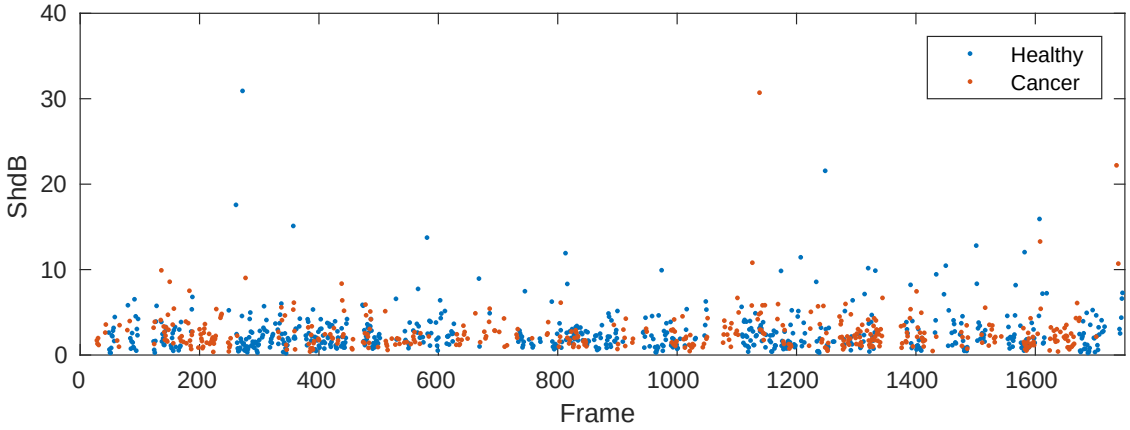


Figure 2.8: The shimmer measurement ShdB for each frame over two full recordings. Both recordings are from the same patient, one before surgery and one after the patient have recovered.

In this section, the different tools used to analyze and classify the data in the feature space will be presented. As the features presented above have been computed they are arranged in a vector $\mathbf{x}_n$, $n \in [1, N]$ for each recording, where N is the number of recordings. The data from all recordings are collected into a matrix $\mathbf{X} = [\mathbf{x}_1 \quad \mathbf{x}_2 \cdots \mathbf{x}_N]^T$. To be able to analyze this data, all features are normalized to be zero mean and have unit variance. This is needed since the Principal Component Analysis presented below seeks to find subspace such that variance is maximized, for this to work properly the variance must be normalized.

2.3.1 Dimensionality reduction and robust handling

Principal Component Analysis - PCA

PCA is an often used tool for data visualization and dimensionality reduction. The main idea is to take D-dimensional data and project it onto a M-dimensional subspace, such that the variance of the data is maximized in the subspace. The idea is that using PCA, one can reduce the dimensionality and only use the components which describe the most of the variability. This tool is used as a way to both reduce the dimensionality of the feature data extracted from the voice recordings and as a way to only use the features were the variation is large.

Consider a set of data points $\mathbf{X}$ with covariance matrix $\mathbf{S}$ in the original D-dimensional space. Now suppose M=1, meaning the subspace on which we wish to project is spanned by a single unit length basis vector $\mathbf{u}_1$. The squared length of the projection of $\mathbf{S}$ is written as $\mathbf{u}_1^{\mathrm{T}}\mathbf{S}\mathbf{u}_1$. By maximizing the length of the projection the direction which corresponds to the maximum variance is found,

$$\begin{aligned} \underset{\mathbf{u}_1}{\text{maximize}} \quad & \mathbf{u}_1^{\mathrm{T}}\mathbf{S}\mathbf{u}_1 \\ \text{subject to} \quad & \mathbf{u}_1^{\mathrm{T}}\mathbf{u}_1 = 1. \end{aligned} \tag{2.12}$$

Using Lagrange multiplier [21], 2.12 may be reformulated as

$$\underset{\mathbf{u}_1}{\text{maximize}} \quad f = \mathbf{u}_1^{\mathrm{T}}\mathbf{S}\mathbf{u}_1 - \lambda_1(1 - \mathbf{u}_1^{\mathrm{T}}\mathbf{u}_1). \tag{2.13}$$

By setting $\dfrac{\partial f}{\partial \mathbf{u}_1} = 0$ and solving for $\mathbf{u}_1$,

$$\mathbf{S}\mathbf{u}_1 = \lambda_1\mathbf{u}_1. \tag{2.14}$$

Multiplying both sides with $\mathbf{u}_1^{\mathrm{T}}$ from the left and making use of the fact that $\mathbf{u}_1$ is a unit vector,

$$\mathbf{u}_1^{\mathrm{T}}\mathbf{S}\mathbf{u}_1 = \lambda_1. \tag{2.15}$$

This means that the projected variance is an eigenvalue to the covariance matrix $\mathbf{S}$, implying that for it to be maximized λ_1 must be the largest eigenvalue of $\mathbf{S}$. Therefore $\mathbf{u}_1$ must be the corresponding eigenvector, or in this context, the first principal component. To obtain proceeding principal

components one simply choose the proceeding eigenvectors with their corresponding eigenvalues in descending order [21]. Finding the eigenvalues can be done in different ways, a common way is to use singular value decomposition.

Robust Principal Component Analysis

When handling data with outliers the PCA may be insufficient. An alternative to PCA so-called robust PCA [22], used in this work to make the dimensionality reduction more robust. Robust PCA assumes that the data is in the form of

$$\mathbf{M} = \mathbf{L_0} + \mathbf{S_0} \tag{2.16}$$

where $\mathbf{L_0}$ is low rank and $\mathbf{S_0}$ is a sparse matrix. It is shown in [22] that finding $\mathbf{L_0}$ and $\mathbf{S_0}$ corresponds to solving the convex optimization problem

$$\begin{aligned} &\text{minimize} \quad ||\mathbf{L}||_* + ||\mathbf{S}||_1 \\ &\text{subject to} \quad \mathbf{L} + \mathbf{S} = \mathbf{M} \end{aligned} \tag{2.17}$$

where $||\mathbf{L}||_*$ denotes the nuclear norm which is defined as the sum of the singular values of $\mathbf{L}$ $\sum_i \sigma_i(\mathbf{L})$ and $||\cdot||_1$ the L_1 norm. This formulation, so-called principal component pursuit, may be solved efficiently by using, for instance, the technique of Alternating Direction of Multipliers (ADMM), in combination with augmented Lagrange multipliers (ALM).

2.3.2 Distribution tests

After having extracted the chosen features from the voice recordings and reduced the dimensionality in a robust way, the different class distributions are tested. It is of great interest to check if the distribution for the pathological and healthy class respectively, are significantly different from each other. This is done by using two tests, one to check if the covariance matrices are significantly different and the other to test if the means are significantly different. The idea is if one is not able to conclude that the distributions are different, there is no point in moving forward and construct a classifier.

Box's M-test

Given the task of determining whether two p-dimensional distributions are significantly different the Box's M-test, introduced in [23], tests if the covariance matrices are significantly different. For example, in figure 2.9 Box's M-test returns positive for the middle and leftmost figure. Box's M makes use of the within class covariances $\mathbf{S}_i$ to form the test statistic

$$M = (N-k)\ln|\mathbf{S}| - \sum_{i=1}^{k}(n_i - 1)\ln|\mathbf{S}_i|, \tag{2.18}$$

where

$$\mathbf{S} = \frac{1}{N-k}\sum_{i=1}^{k}(n_i - 1)\mathbf{S}_i \tag{2.19}$$

where n_i and N is the number of data points for each class and in total, respectively [23], $|\mathbf{S}|$ denotes the determinant of the matrix, and k is the number of classes, in this work $k = 2$. If the two class covariances are identical, M will be zero, otherwise greater than zero. By testing if M is significantly different from zero, one can determine if the covariances are significantly different.

M is approximately distributed as

$$\chi_v^2 = M(1 - C), \tag{2.20}$$

where the degrees of freedom

$$v = \frac{p(p+1)}{2}, \tag{2.21}$$

and

$$C = \frac{2p^2 + 3p - 1}{6(p+1)}\Big(\frac{1}{n_1 - 1} + \frac{1}{n_2 - 1} - \frac{1}{N-2}\Big) \quad . \tag{2.22}$$

The test is then conducted on a 5 % significance level, to determine whether the distributions are different or not. This work uses a slightly modified version of the implementation provided in [24].

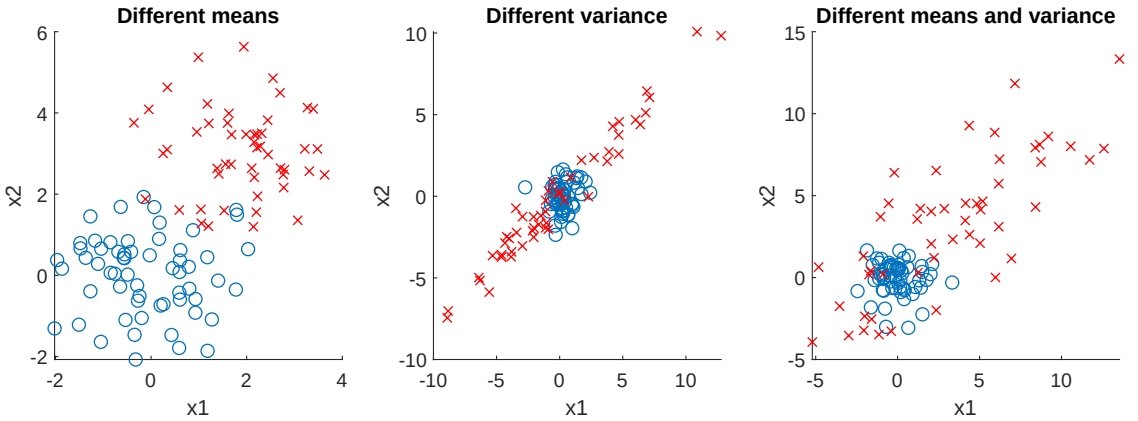


Figure 2.9: Examples of distribution types to test. The difference of the covariance and mean for the two different classes is tested. The Box's M-test is used to determine if the covariances are different (middle and left figure) and the Modified Nel and Van der Merwe test determines if the means are different(left and right figure).

Modified Nel and Van der Merwe test

In addition to testing the difference of the covariance matrices, the difference between the means of two p-dimensional distributions with N_i samples is tested using the modified Nel and Van der Merwe (MNV) test [25]. Meaning MNV is useful for detecting the left- and rightmost distributions in figure 2.9. Using the multivariate mean of the two classes $\bar{\mathbf{X}}_\mathbf{1}$, $\bar{\mathbf{X}}_\mathbf{2}$ and the sample covariances

$$\tilde{\mathbf{S}}_\mathbf{i} = \frac{1}{N_i(N_i-1)} \sum_{j=1}^{N_i} (\mathbf{X}_{ij} - \bar{\mathbf{X}}_\mathbf{i})(\mathbf{X}_{ij} - \bar{\mathbf{X}}_\mathbf{i})', \quad i = 1, 2 \tag{2.23}$$

a test statistic

$$T^2 = (\bar{\mathbf{X}}_\mathbf{1} - \bar{\mathbf{X}}_\mathbf{2})' \tilde{\mathbf{S}}^{-1} (\bar{\mathbf{X}}_\mathbf{1} - \bar{\mathbf{X}}_\mathbf{2}) \tag{2.24}$$

where $\tilde{\mathbf{S}} = \tilde{\mathbf{S}}_\mathbf{1} + \tilde{\mathbf{S}}_\mathbf{2}$ is formed.

It is shown in [25] that T^2 approximately follows an F-distribution multiplied by a constant

$$T^2 \sim vpF_{p,v-p+1}/(v-p+1) \tag{2.25}$$

where v is the approximate degrees of freedom for the distribution of $\tilde{\mathbf{S}}$. It depends on p and the covariances $\tilde{\mathbf{S}}$, $\tilde{\mathbf{S}}_\mathbf{1}$, and $\tilde{\mathbf{S}}_\mathbf{2}$ and is not presented in detail here.

By testing the computed value of T^2 against the 95th percentile of the F-distribution it is determined if the means are significantly different on a confidence level of 0.05.

2.3.3 Classification

After pre-processing the data and testing for different distributions, the next goal is to be able to classify the data, as either healthy or pathological.

Support vector machine - SVM

The classifier primarily used in this work is a Support Vector Machine (SVM).

Considering a data point $\mathbf{x}$ corresponding to the features of one recording. Each point $\mathbf{x}$ has a corresponding label t which is either +1 or -1. The classification task can then be formulated on the form

$$y(\mathbf{x}) = \mathbf{w}^T \phi(\mathbf{x}) + b \tag{2.26}$$

where $\phi(\mathbf{x})$ is a fixed transformation, for a linear kernel SVM $\phi(\mathbf{x}) = \mathbf{x}$, $\mathbf{w}$ is a weighting vector and b a bias term. Using these terms, a prediction $y(\mathbf{x})$ of the true label t is computed. The basic assumption of the SVM is that the two classes are linearly separable, if this is the case there exists a $\mathbf{w}$ and b such that $y(\mathbf{x}_n) < 0$ for all $\mathbf{x}_n$ with corresponding $t_n = -1$ and similarly $y(\mathbf{x}_n) > 0$ for $t_n = 1$. Or equivalently $t_n y(\mathbf{x}_n) > 0$ for all $\mathbf{x}_n$. Many such choices of $\mathbf{w}$ and b, describing a

separating hyperplane may exist and the idea of the SVM is to choose the one that maximizes the *margin*, that is the perpendicular distance from the hyperplane to the closest data point. These data points that lie on the margin are known as support vectors. It can be shown that finding the hyperplane that maximizes the margin is equivalent to solving

$$\begin{aligned} \underset{\mathbf{w},b}{\arg\min} \quad & \frac{1}{2}||\mathbf{w}||^2 \\ \text{subject to} \quad & t_n(\mathbf{w}^T + b) \geq 1 \end{aligned} \tag{2.27}$$

In many cases however, the class distributions are overlapping. This is handled when training the SVM by penalizing data points that are on the wrong side of the margin. This is done by introducing a slack variable ξ_n for each data point $\mathbf{x}_n$. For data points on the correct side of the margin $\xi_n = 0$, for data points on the wrong side of the margin $\xi_n = |t_n - y(\mathbf{x}_n)|$. Implying that for points on the wrong side of the margin but on the correct side of the decision boundary $0 < \xi_n < 1$, and $\xi_n \geq 1$ for points on or on the wrong side of the decision boundary. The optimization problem is then reformulated as

$$\begin{aligned} \underset{\mathbf{w},b}{\arg\min} \quad & C\sum_{n=1}^{N}\xi_n + \frac{1}{2}||\mathbf{w}||^2 \\ \text{subject to} \quad & t_n(\mathbf{w}^T + b) \geq 1 - \xi_n, \end{aligned} \tag{2.28}$$

where the parameter C determines how much emphasis should be put on wrongly classified points, here $C = 1$ [21].

Other classifiers

Besides SVM other types of classifiers were experimented with as well. Mainly shallow neural networks and k nearest neighbour. The results were not better than that for the SVM, due to the speed of training a SVM in combination with its theoretical properties the author chose to continue working primarily with SVM as classifier and therefore neither theory or results for other classifiers is presented here.

Results

3.1 Stationary data

For each frame, of which there typically are between 20-80 for each recording, the features presented earlier were computed, the median of each feature over the entire recording was computed, as well as the median of the frame by frame difference for each recording. In figure 3.10 the distributions of the median for three different features as well as the median of the frame by frame difference Δ. The distributions for the different labels are different, however overlapping. This shows us that while there may very well be useful information in the coefficient. It is not possible to make a distinction by simply setting a threshold value. Only three feature distributions are shown here, although many others exhibit the same type of characteristics. A shared trait for these three features is that it is common for the absolute value of the frame by frame difference to be larger for Laryngitis patients. Therefore, to simplify for a classification algorithm to assign weights, the absolute value of the frame by frame difference is saved as a feature.

The median over the entire recording as well as the median of the frame by frame difference is displayed in 3.10. In table 3.2 the corresponding correlations are shown.

A natural question is if the distributions for the features and the distribution for the frame by frame difference is correlated. If the correlation is high, it would be suboptimal to use both as features in classifying voices. The correlations are presented in table 3.2 and it can be seen that the correlations are low, indicating that using both the parameter and the difference provides more information than simply using one of them.

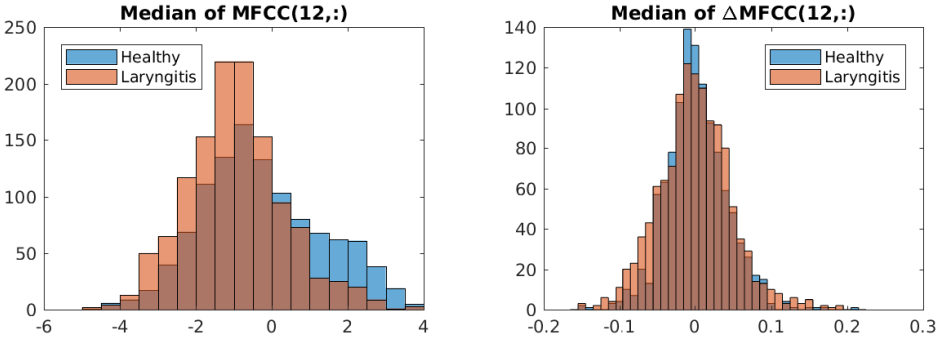
A more thorough investigation of the correlation between the median, standard deviation and frame by frame difference for each parameter shows that for the Saarbrücken voice data, out of 23 different parameters, all but six have a correlation lower than 0.5 and all have a correlation lower than 0.75. For the non-stationary data, the corresponding numbers are 7 and 0, respectively.

In order to further analyze the data extracted from the recordings, the data was normalized to zero mean and unit variance in order to make different features comparable to each other. All recordings were at least one of the features were missing were removed, for instance jitter and shimmer were commonly missing if the correct number of peaks in the time domain was not found. It was found that out of all recordings missing values, 76 % of them were from persons with laryngitis, suggesting that this is perhaps an indication that the voice quality sometimes is unsurprisingly bad for laryngitis patients.

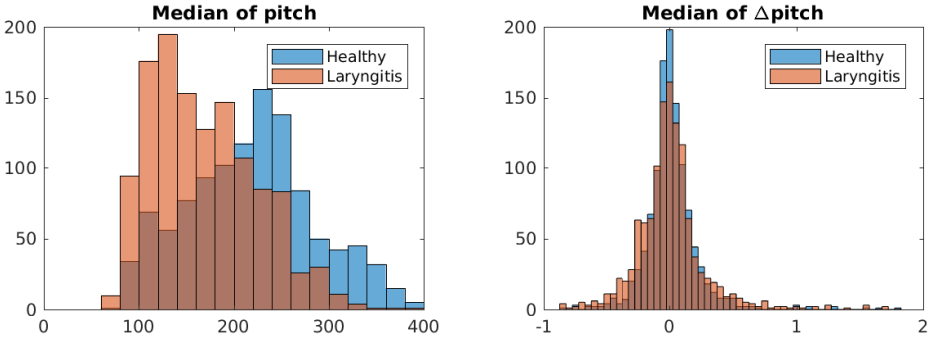
The two first PCA vectors of the feature values is shown in figure 3.11. Ideally, we would have seen two clusters, one red and one blue. That would mean that the healthy voices and laryngitis voices would be linearly separable in this subspace. That is, however, not the case, the overlap of the clusters indicate that there is not a simple way of classifying the voices. An issue with the distribution shown is that there seems to be a rather large number of outliers. To combat this, the robust PCA is used to extract $\mathbf{L}$, as defined in equation (2.16), out of the normalized data. The matrix $\mathbf{L}$ projected onto its two first principal components is shown in figure 3.12. The effect of the robust PCA is shown in the figure as the outliers present when using PCA are now gone.

After reducing the dimensionality by using robust PCA, a natural next step is to analyze the distributions of the data in the space of the principal components. In figure 3.14, a normal probability plot of the data projected onto the first principal component is shown. By viewing the figure, it is very reasonable to question the assumption soon to be made that the data is normal distributed. The data projected onto the following principal components exhibits a similar behaviour, and for most the assumption of normality is a better assumption than for the first component.

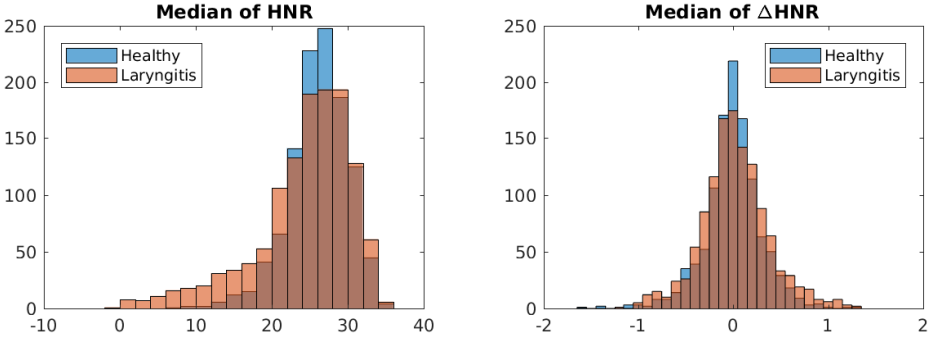
To test if the means for the healthy and pathological classes are significantly different from



(a) Distribution of the 12th MFCC.



(b) Distribution of the pitch.



(c) Distribution of the HNR.

Figure 3.10: Distributions for a few different features on the stationary data. One the left the distributions of the median of the feature over an entire recording is shown. On the right the median of the frame by frame difference is shown.

each other, they are tested using the modified Nel and Van der Merwe test [25]. As the rank of $\mathbf{L}$ is 39, the distribution of the two classes were tested on the 39 first principal components. In the complete space of the 39 principal components, the distributions were found to be significantly different. Studying the univariate distribution in each direction, it was found that it was different in 19 components, including the first four which are explaining 85 % of the variability. When discarding all components explaining less than 1 %, 7 components remained, namely components 1-4 and 5-8, these 7 components explain 91 % of the variability.

Having reduced the data from 69 dimensions to 7 and concluding that the two class distributions are significantly different, the next step is to construct a classifier. The data was divided into 10 parts, and an SVM was trained 10 different times, each time using 9 parts as training data and 1 part as validation data. Using this so called 10-fold cross validation technique, some information of the performance of the classifier is obtained. The classifier assigns a value for classification to each data point. By varying which threshold to use for classification, the different correct classification rates as well as misclassification rates varies. For instance, by setting a very high threshold for classification, the true negative rate will be high, as well as the false negative rate. Determining an appropriate threshold is an important task where the characteristics of the test

Table 3.2: Correlation between the median of a parameter over a voice signal and the median of its corresponding frame by frame difference Δ.

Feature	Correlation
$MFCC_{12}$	0.006
pitch	0.047
HNR	0.063

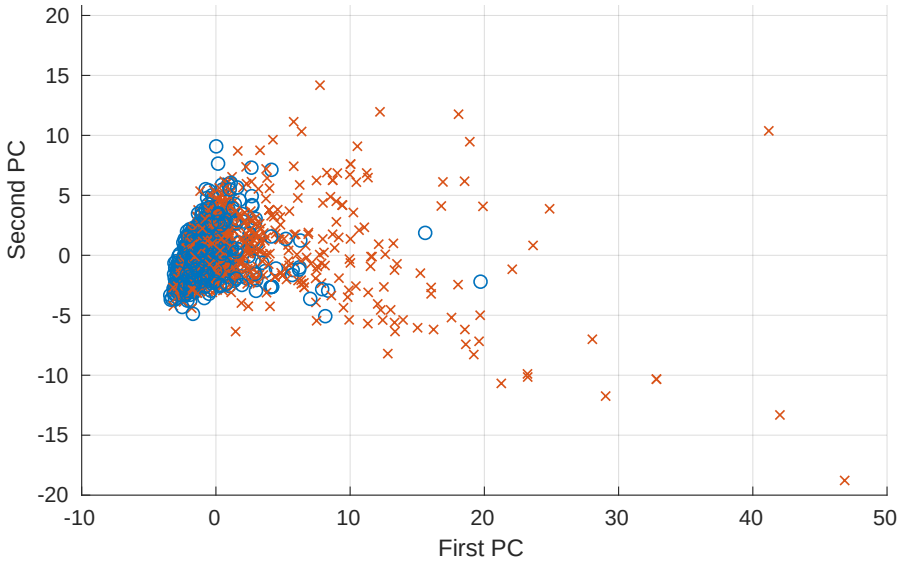


Figure 3.11: Principal component analysis of the feature matrix extracted from the stationary voice data. The data is normalized to zero mean and unit variance. Red x's denote recordings from persons with Laryngitis and blue o's show healthy persons. The first three components shown here explain 31 % of the variability. The first 54 components are needed to explain 99 % of the variability.

needs to be considered. A first approach is to simply go for the highest possible accuracy, the task may however, be more complex. For instance whether a false negative or false positive result is more detrimental. In this case, a false positive result may induce unnecessary anxiety whereas a false negative will cause the test to miss a reoccurring tumour. These types of questions are out of the scope of this thesis, they are however, important to consider.

In figure 3.15, a receiver operating characteristics (ROC) curve where the true positive rate is plotted over the false positive rate is displayed. The curve is shown for both training and validation set for each of the 10 cross validations. When lowering the threshold the true positive rate increases, however, so does the false positive rate. The baseline for classification is the grey dashed line where the true positive rate increases at the same rate as the false positive rate. A random classifier would be expected to perform in this manner. The area under the curve (AUC) is a commonly used measurement to evaluate a classifiers performance, where the dashed grey line has an area under the curve of 0.5. In table 3.4, the mean as well as standard deviation of the AUC measurements on the training and validation data is displayed. Also the maximum accuracy on the training set as well as the accuracy on the validation set using the "best" (in the sense of maximum accuracy) threshold from the training data. The performance of the classifier is while being clearly better than the baseline of a random classifier, not very exciting in comparison with some earlier research presented earlier. For instance [9] obtained over 80% accuracy using similar features.

3.2 Non-stationary data

Using the same technique for feature extraction as for the stationary voice data and reducing the dimensionality using PCA results in the plot shown in figure 3.16. The figure shows that separating data labeled pathological and healthy respectively, is not trivial. The two classes are not linearly

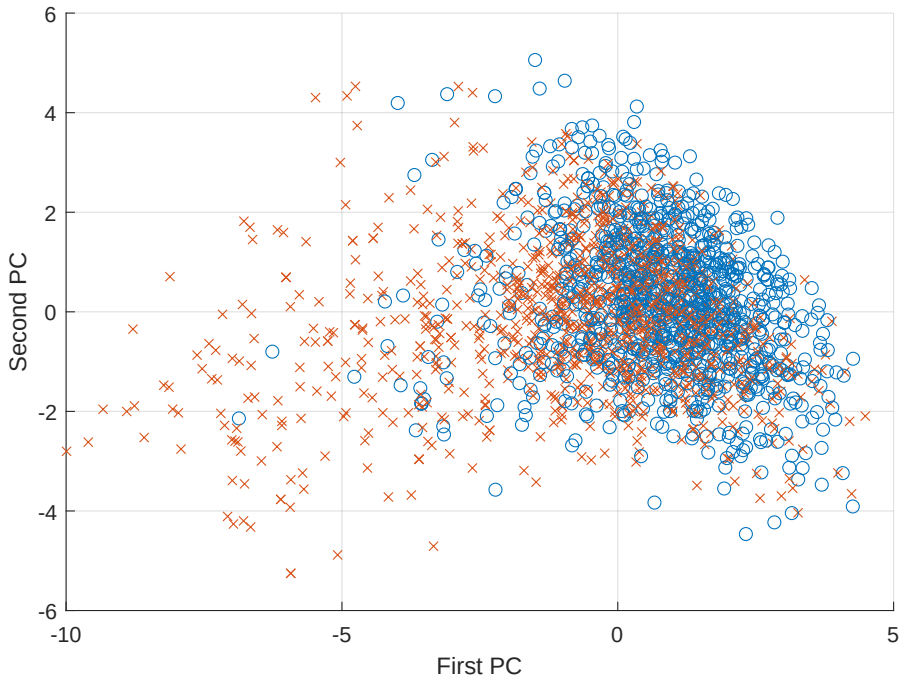


Figure 3.12: Robust principal component analysis of the feature matrix extracted from the stationary voice data. The data is normalized to zero mean and unit variance. Red x's denote recordings from persons with Laryngitis and blue o's show healthy persons. The first two components shown here explain 59 % of the variability. To explain 99 % of the variablity the first 16 components are needed.

Table 3.3: Confusion matrix when the threshold for classification is -0.47, 0.005 and 1.41 respectively. Corresponding accuracy on the validation set is 0.68 ± 0.029, 0.73 ± 0.030, and 0.64 ± 0.033.

True positive	False positive
False negative	True negative

91	59
14	52

79	38
26	73

79	38
26	73

separable in the space spanned by the first three PCs.

An important difference between this set of recordings as compared to the Saarbrücken data set is that this set contains data with belonging to both classes from the same person. This comes with the benefit that the natural variations between people are eliminated. Among the 24 patients there are four from whom there are six recordings or more. In figures 3.17 - 3.20, the data for single patients is shown. In all cases except for figure 3.19, the two classes are linearly separable in the space of the first three PCs.

A natural question is how well a classifier trained on the stationary data generalizes to the non-stationary data. When this was investigated, it was concluded that the classifier generalized poorly and the prediction results were no better than a random guess.

To be able to separate the classes, the first question that arise is whether the distributions are significantly different. By performing the modified Nel and Van der Merwe test, both the univariate distribution for each of the principal components, as well as multivariate test adding on one component at a time, components in which the distributions are separate are found. On a significance level of 0.05 it was found that the covariance matrices is significantly different for the two classes. However the modified Nel and Van der Merwe found that the mean of the distribution are not significantly different. To pass the multivariate Box's M-test, the first five or more principal components were needed for the distributions to be deemed significantly different.

In the univariate case, the data was found to have different means in the 5:th and 25:th component. The 25:th component is of less interest as it explains less than 0.1 % of the variability. The 5:th PC however explains 4.7 %. In figure 3.21, the distribution of the data in the 5:th component is shown. As it is assumed for the tests that the data is normal distributed, the a normality plot is shown in figure A.1. The normality assumption is not unreasonable. From figure 3.21, it is easily seen that classification by threshold will not yield very satisfactory results; in fact the maximum

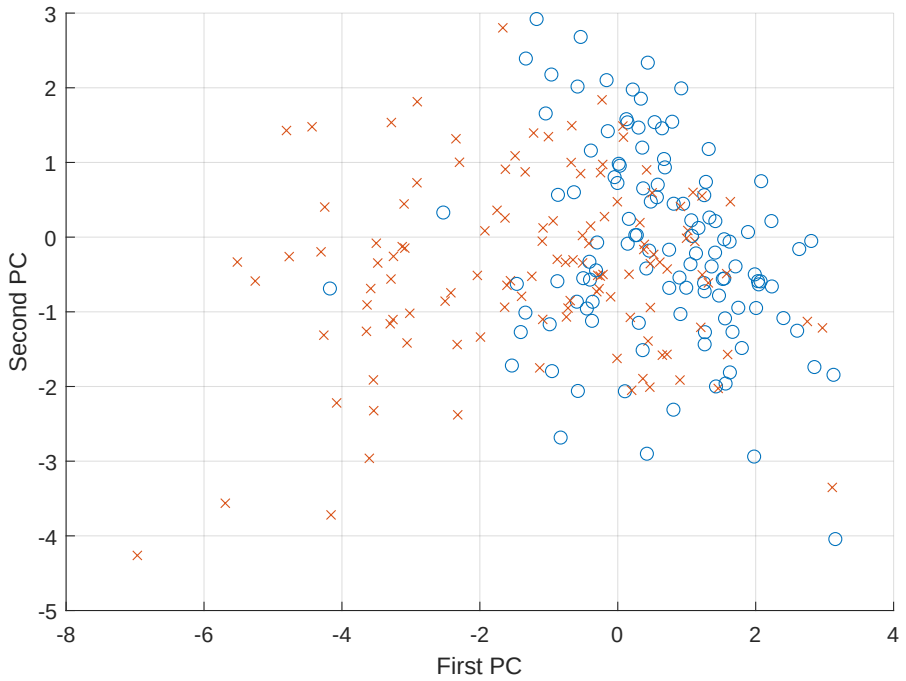


Figure 3.13: The same as figure 3.12, with the difference that only one recording per person is used. 57 % of the variability is explained by the first two components. Here, 22 components are needed to explain 99 % of the variability.

Table 3.4: Statistics of an SVM classifier using 10-fold cross validation.

Measurement	**Mean**	**Standard deviation**
AUC of training data	0.8051	0.0022
AUC of validation data	0.8015	0.0194
Maximum accuracy on training set	0.7349	0.0029
Corresponding accuracy on validation set	0.7292	0.0270

accuracy by threshold is 64 %. The ROC curve of the classification is shown in figure A.2, as well as the corresponding confusion matrix in table A.1.

Given that there is recordings from the same patient belonging to both classes in this data set simplifies the problem slightly. As it is possible to look at a single patient and therefore eliminate the great variability in different people's voices. For the four patients from whom there are six or more available recordings, the same type of analysis was made. The issue, however, is the small number of data points. Given that the minority class often only contains two data points, it is impossible to determine whether a normal distribution is a good assumption. The multivariate modified Nel and Van der Merwe test deemed that the distributions were not significantly different. For the univariate case, however, the distributions were different in some of the components. For two of the patients, the distributions were different in the 6th principal component. In figure 3.22, the value of the data projected onto the 6th principal component is shown as a function of the patients state over time. For these two patients, the sixth principal component distinguish the two classes well. Unfortunately, this is not the case for the two other patients considered here. For one of the patients the distributions were different in the third principal component, the data for this patient is shown in figure 3.23. For the fourth patient, there was no univariate distribution in which the classes were different. However, for the multivariate distributions of PCs 1-5 and 1-6, the classes were significantly different. Before concluding anything further it is important to remind ourselves of the very crude assumptions previously made. It seems as if the components 3, 5, and 6 may be of slightly more use when attempting to classify the voices of single persons. Also, as mentioned earlier, when making use of the entire data set, the fifth principal component was of interest. In table 3.5, the 5 most weighted features and their correspinding weights in PCs 3, 5, and 6 is shown. What is remarkable is that all features are related to the MFCCs.

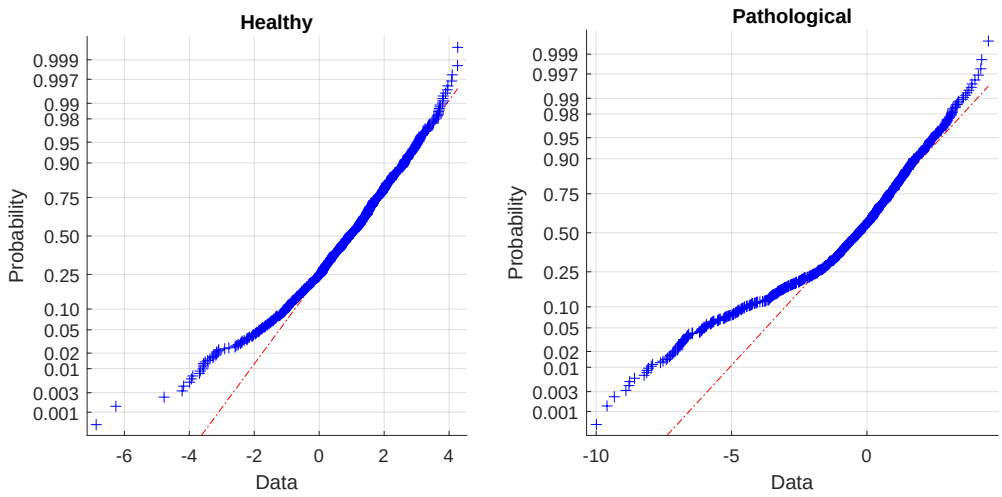


Figure 3.14: Normal probability plot of the data projected onto the first principal component.

Table 3.5: Most relevant features

PC 3	**PC 5**	**PC 6**
Δ $MFCC_9$ 0.32	std $MFCC_9$ 0.35	std $MFCC_6$ 0.35
Δ $MFCC_{12}$ 0.31	median $MFCC_6$ 0.35	median $MFCC_4$ -0.31
Δ $MFCC_5$ 0.29	median $MFCC_5$ -0.28	std $MFCC_{11}$ 0.27
Δ $MFCC_{10}$ 0.26	median $MFCC_{11}$ 0.25	median $MFCC_9$ 0.25
Δ $MFCC_6$ 0.24	median $MFCC_1$ 0.24	median $MFCC_2$ -0.22

It would be natural to make a classifier for this data as well as for the stationary data in the previous section. However, given that the distributions are not different in more than one component, one could not expect to get better accuracy than simply using a threshold in one dimension as presented earlier. In the case of only using data from a single person, there are too few data points to try and build a classifier. Hopefully, a larger amount and more regular data from the same patient will be available in the future. If this would be the case, it seems as if the MFCCs would be a good place to start.

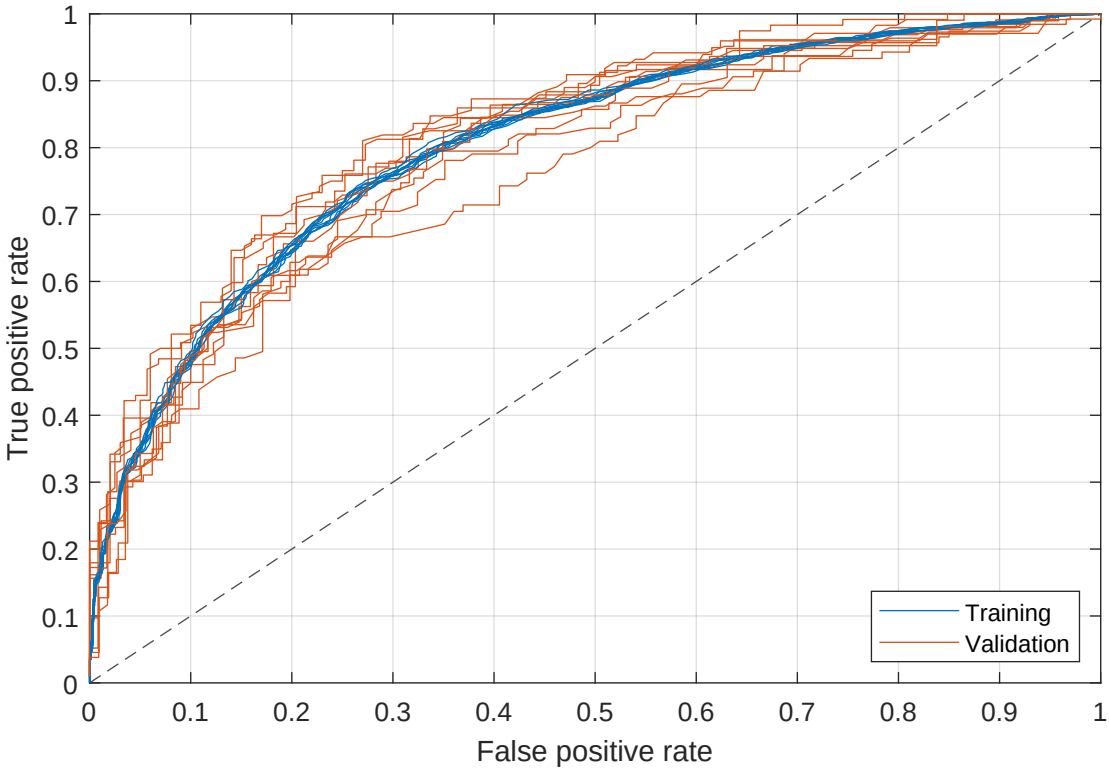


Figure 3.15: ROC curve plotting the true positive rate over the false positive rate for 10-fold cross validation of an SVM classifier.

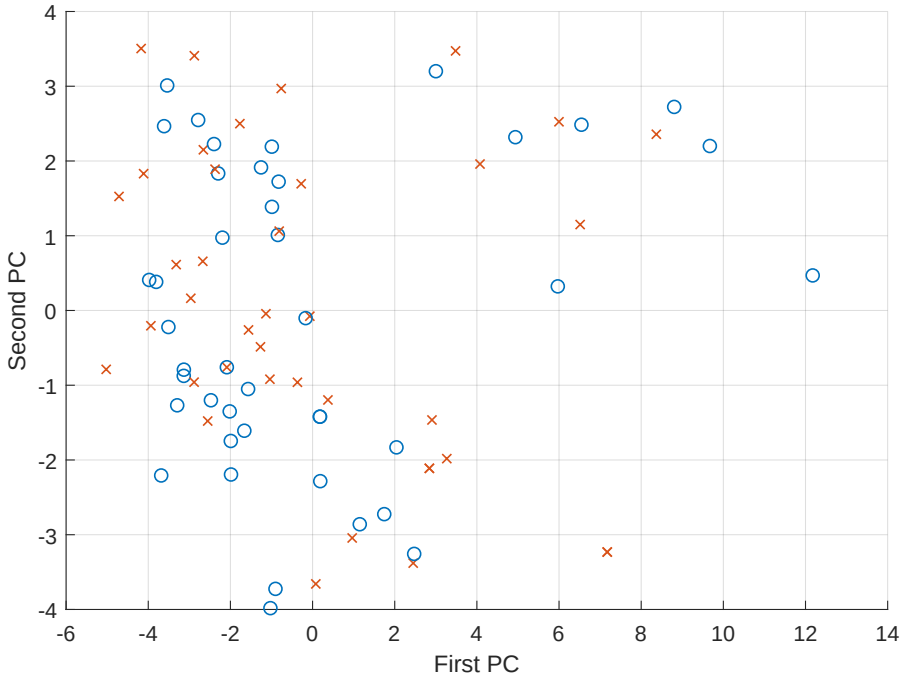


Figure 3.16: PCA of data computed from 82 recordings from in total 24 patients. Red x's denote recordings from when patients had a condition and blue o's denote the recordings from when patients where healthy. The first three components explain 41 % of the variability in the data.

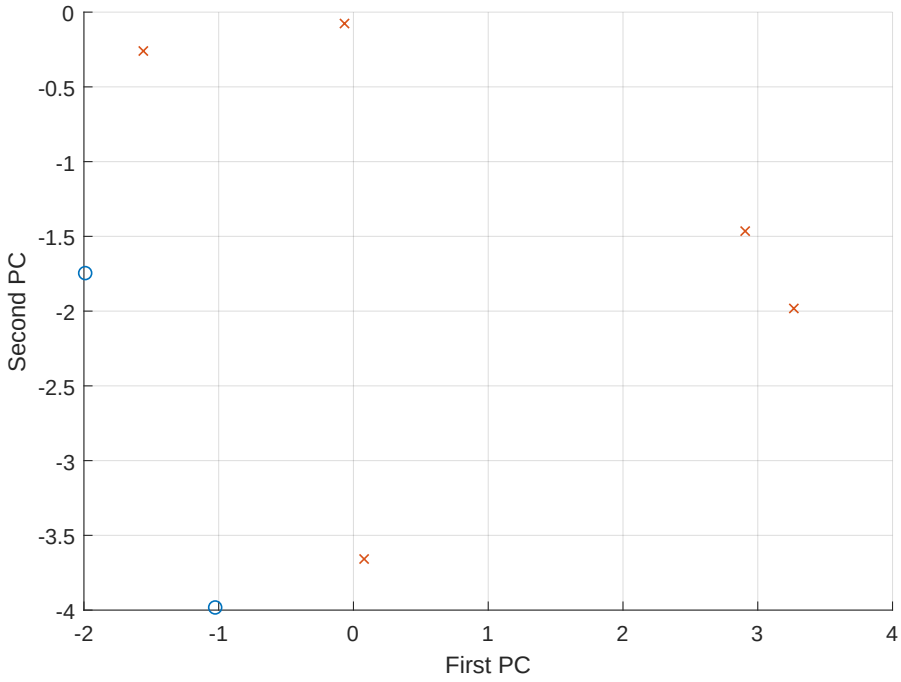


Figure 3.17: PCA of the same data as in figure 3.16 showing only the points from a single patient.

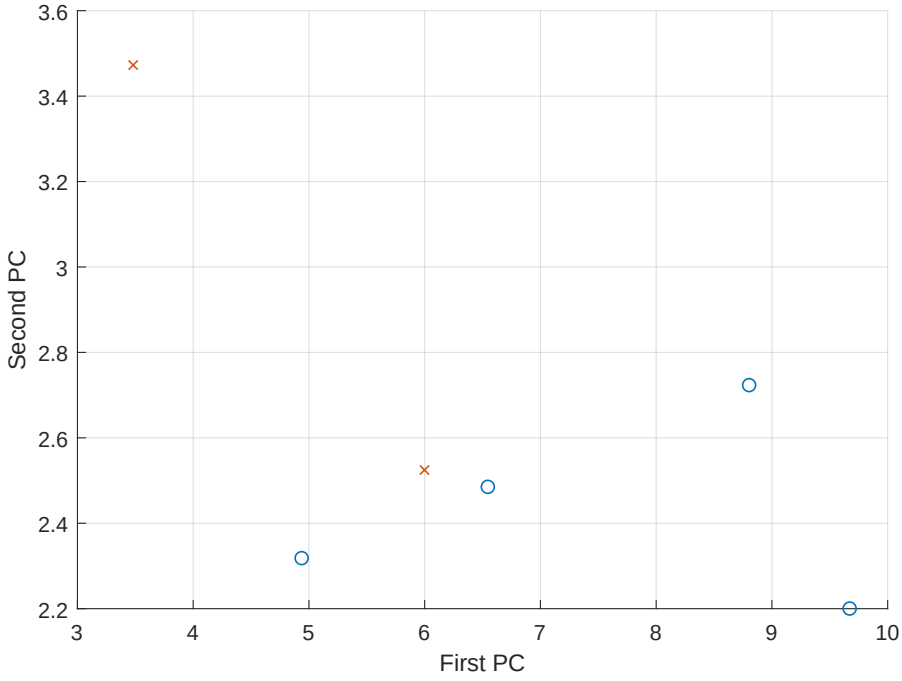


Figure 3.18: PCA of the same data as in figure 3.16 showing only the points from a single patient.

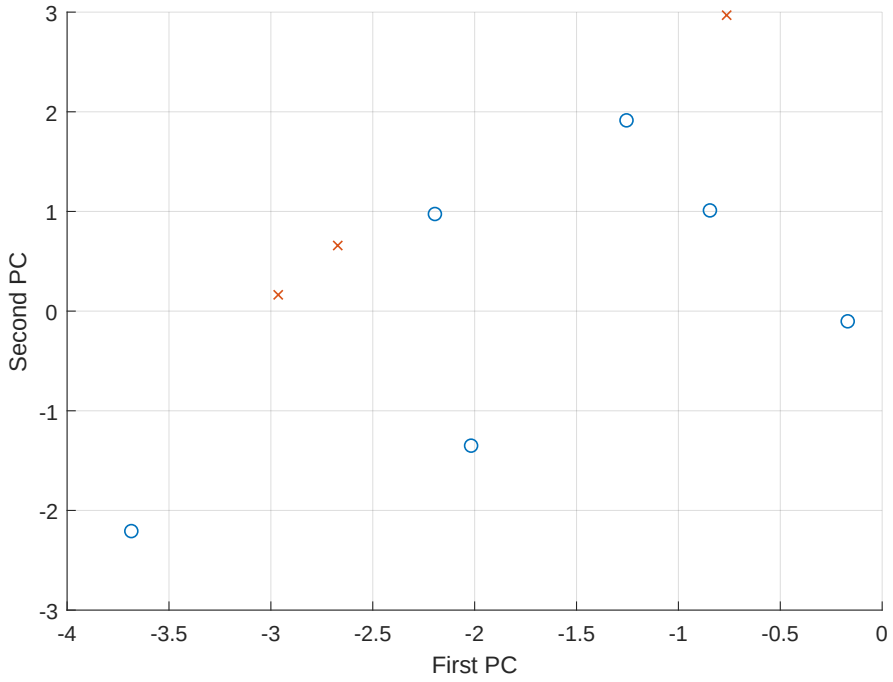


Figure 3.19: PCA of the same data as in figure 3.16 showing only the points from a single patient.

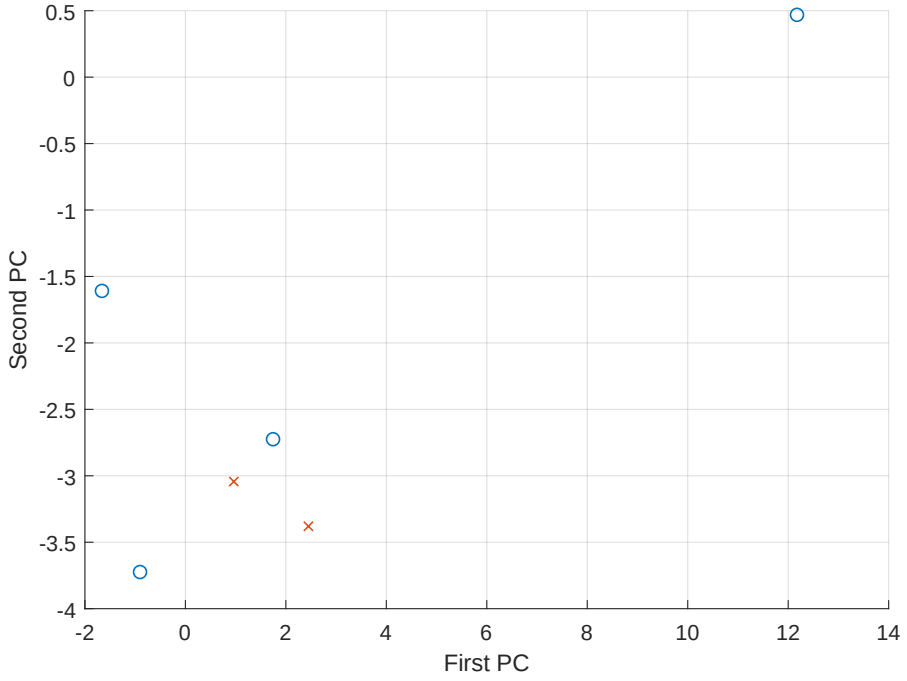


Figure 3.20: PCA of the same data as in figure 3.16 showing only the points from a single patient.

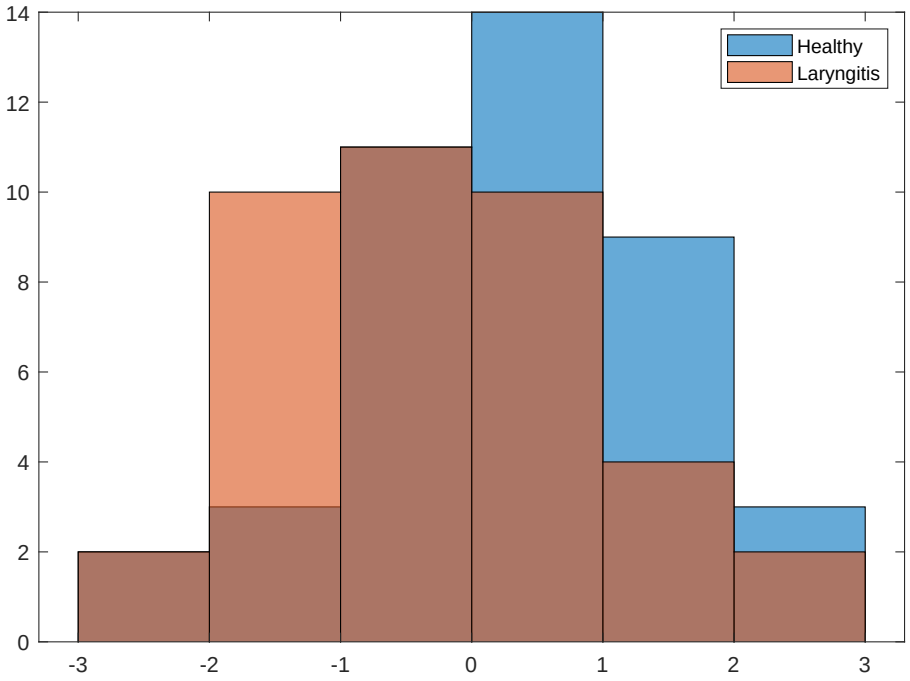
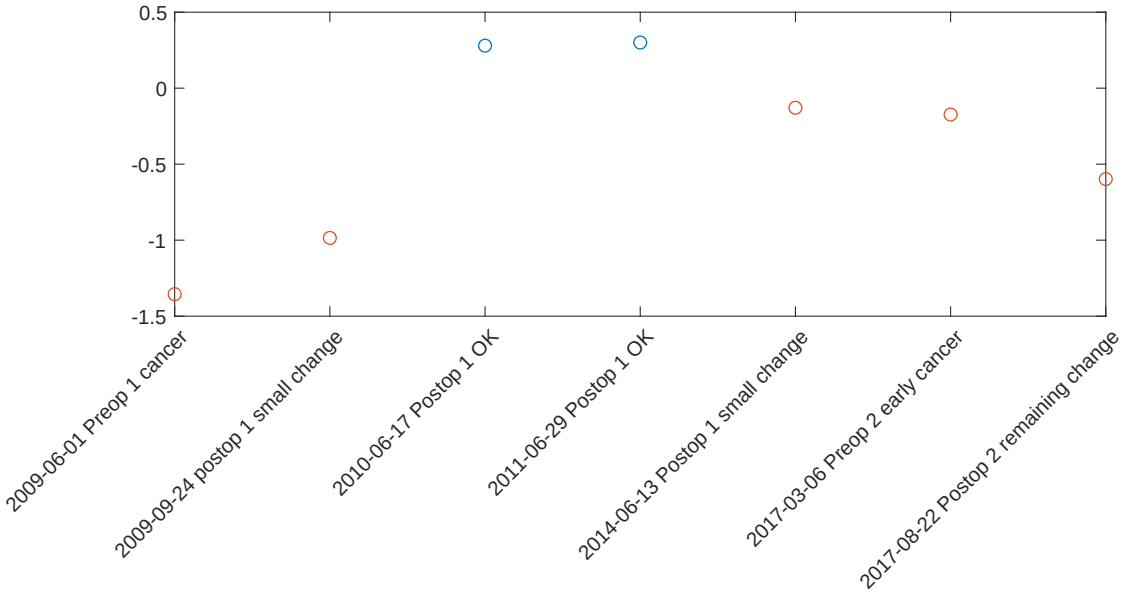
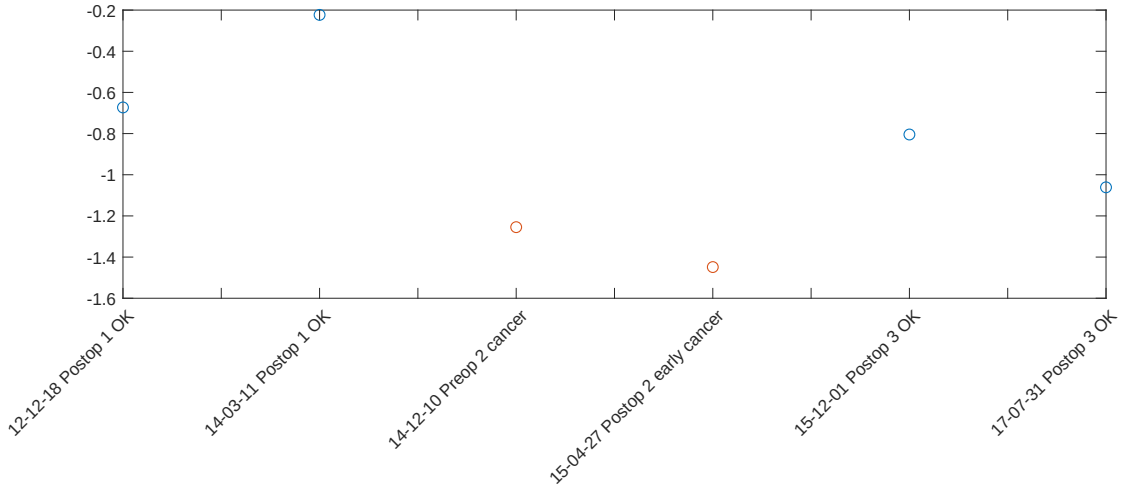


Figure 3.21: Fifth principal component



(a) Patient 2



(b) Patient 4

Figure 3.22: Value of the 6th principal component over time for two different patients.

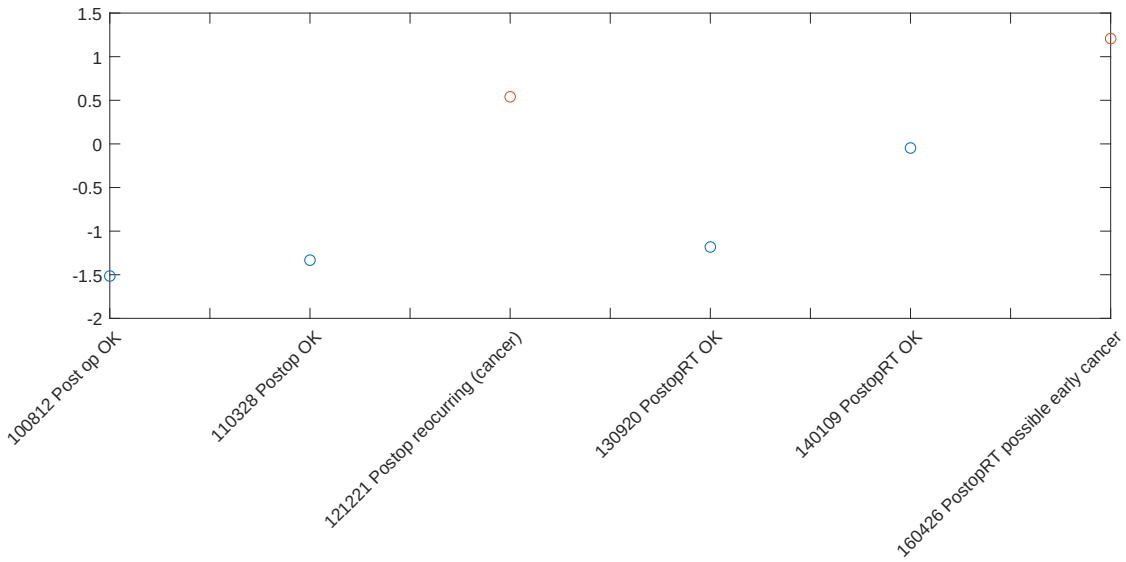


Figure 3.23: Value of the 3rd principal component over time for one patient.

Discussion and Suggested further research

An obvious conclusion from this thesis is that the classification of stationary voice data is sub par in comparison to some of the results presented in earlier research. This is surprising due to the similarity of both the framing procedure, features used and classification method to some earlier research that obtained more satisfactory results. Small variations in implementation differences may have caused these poor results. For instance, when computing jitter and shimmer, it is necessary to find the correct peaks, the method of finding the peaks may vary. For the MFCCs though, the variation in implemention should be smaller, due to its non parametric nature. However, the number of filter banks used as well as number of coefficients kept can still vary. Another explanation may be that more advanced classification methods are needed. For instance, in [8] the CCR is approximately 64 % with an AUC of around 0.8 using only single vowels from the Saarbrücken voice database. Comparable to the results obtained here. It is when they combine the classifiers trained on different data that an increase in performance is obtained and they get a CCR of 79 % and AUC of 0.89. This is still not as high as the best classification rates above 90 % but it shows that it is possible to obtain better classifications by combining data. Other types of classifiers such as Gaussian Mixture Models (GMM) may also increase performance. Another approach that could very well be of interest to experiment with is to use classification by committee where multiple classifiers vote to classify data. In short there is a lot to be tested on the classification of the data as this project has mostly focused on the extraction of features from the voice data and less on the classification part. A recommendation for a continuation of this work would also be to use more features, primarily linear prediction coefficients and their cosine transform.

This project aimed to reduce each recording to a feature vector, this to provide a rather simple and comprehensive representation of the state of each recording. This proved to be rather inflexible, another possible angle to explore is to construct multiple different feature vector, one for each type of feature. For instance, one for perturbation measurements (jitter, shimmer), another for MFCCs and a third one for the Δ coefficients. Then train different classifiers for each feature vector and possibly combine them is some sort of committee. This would probably yield a better flexibility and allow to easier remove and add features, with the downside of added complexity.

One of the most interesting parts of this project is the second data set where there are recordings available from the same patients over time. The limitation here, however is the small number of data points as well as how far apart in time they are. In addition the recordings are of short stories which have proven much more difficult to analyze as opposed to vowels. Hopefully more frequent recordings as well as extended vowel pronunciations will be available soon. The author is convinced that such a data set would provide great promise for automatic detection of reoccurring cancer.

Bibliography

[1] V. Mittal and Y. Sharma, "Voice Parameter Analysis for the disease detection," *IOSR Journal of Electronics and Communication Engineering Ver. I*, vol. 9, no. 3, pp. 48–55, 2014. [Online]. Available: www.iosrjournals.org.

[2] Z. Witold Engel, M. Kłaczyński, and W. Wszołek, "A vibroacoustic model of selected human larynx diseases," *International Journal of Occupational Safety and Ergonomics*, vol. 13, no. 4, pp. 367–379, 2007, ISSN: 10803548. DOI: 10.1080/10803548.2007.11105094.

[3] M. Shahbakhti, D. Taherifar, and Z. Zareei, "Combination of PCA and SVM for diagnosis of Parkinson's disease," *2013 2nd International Conference on Advances in Biomedical Engineering, ICABME 2013*, pp. 137–140, 2013. DOI: 10.1109/ICABME.2013.6648866.

[4] A. Tsanas, M. A. Little, P. E. McSharry, and L. O. Ramig, "Accurate telemonitoring of parkinsons disease progression by noninvasive speech tests," *IEEE Transactions on Biomedical Engineering*, vol. 57, no. 4, pp. 884–893, 2010, ISSN: 00189294. DOI: 10.1109/TBME.2009.2036000.

[5] H. F. Wertzner, S. Schreiber, and L. Amaro, "Analysis of fundamental frequency, jitter, shimmer and vocal intensity in children with phonological disorders," *Brazilian journal of otorhinolaryngology*, vol. 71, no. 5, pp. 582–588, 2005, ISSN: 1808-8694. DOI: /S0034-72992005000500007. [Online]. Available: http://dx.doi.org/10.1016/S1808-8694(15)31261-1.

[6] J. P. Teixeira and A. Gonçalves, "Algorithm for Jitter and Shimmer Measurement in Pathologic Voices," *Procedia Computer Science*, vol. 100, pp. 271–279, 2016, ISSN: 18770509. DOI: 10.1016/j.procs.2016.09.155.

[7] C. T. Ferrand, "Harmonics-to-Noise Ratio," *Journal of Voice*, vol. 16, no. 4, pp. 480–487, 2002, ISSN: 08921997. DOI: 10.1016/S0892-1997(02)00123-6.

[8] D. Martínez, E. Lleida, A. Ortega, A. Miguel, and J. Villalba, "Voice Pathology Detection on the Saarbrücken Voice Database with Calibration and Fusion of Scores Using MultiFocal Toolkit," pp. 99–109, 2012.

[9] V. Uloza, A. Verikas, M. Bacauskiene, A. Gelzinis, R. Pribuisiene, M. Kaseta, and V. Saferis, "Categorizing normal and pathological voices: Automated and perceptual categorization," *Journal of Voice*, vol. 25, no. 6, pp. 700–708, 2011, ISSN: 08921997. DOI: 10.1016/j.jvoice.2010.04.009. [Online]. Available: http://dx.doi.org/10.1016/j.jvoice.2010.04.009.

[10] G. Muhammad and M. Melhem, "Pathological voice detection and binary classification using MPEG-7 audio features," *Biomedical Signal Processing and Control*, vol. 11, no. 1, pp. 1–9, 2014, ISSN: 17468108. DOI: 10.1016/j.bspc.2014.02.001. [Online]. Available: http://dx.doi.org/10.1016/j.bspc.2014.02.001.

[11] A. Akbari and M. K. Arjmandi, "An efficient voice pathology classification scheme based on applying multi-layer linear discriminant analysis to wavelet packet-based features," *Biomedical Signal Processing and Control*, vol. 10, no. 1, pp. 209–223, 2014, ISSN: 17468108. DOI: 10.1016/j.bspc.2013.11.002. [Online]. Available: http://dx.doi.org/10.1016/j.bspc.2013.11.002.

[12] A. B. Aicha and K. Ezzine, "Cancer Larynx Detection Using Glottal Flow Parameters and Statistical Tools," *2016 International Symposium on Signal, Image, Video and Communications (ISIVC)*, no. 1, pp. 65–70, 2016.

[13] G. Muhammad, M. Alsulaiman, Z. Ali, T. A. Mesallam, M. Farahat, K. H. Malki, A. Al-nasheri, and M. A. Bencherif, "Voice pathology detection using interlaced derivative pattern on glottal source excitation," *Biomedical Signal Processing and Control*, vol. 31, pp. 156–164, 2017, ISSN: 17468108. DOI: 10.1016/j.bspc.2016.08.002. [Online]. Available: http://dx.doi.org/10.1016/j.bspc.2016.08.002.

[14] M. Hariharana, K. Polatb, R. Sindhuc, and S. Yaacoba, "A hybrid expert system approach for telemonitoring of vocal fold pathology," *Applied Soft Computing Journal*, vol. 13, no. 10, pp. 4148–4161, 2013, ISSN: 15684946. DOI: 10.1016/j.asoc.2013.06.004. [Online]. Available: http://dx.doi.org/10.1016/j.asoc.2013.06.004.

[15] A. I. Fontes, P. T. Souza, A. D. Neto, A. D. M. Martins, and L. F. Silveira, "Classification system of pathological voices using correntropy," *Mathematical Problems in Engineering*, vol. 2014, 2014, ISSN: 15635147. DOI: 10.1155/2014/924786.

[16] N. Sáenz-Lechón, J. I. Godino-Llorente, V. Osma-Ruiz, and P. Gómez-Vilda, "Methodological issues in the development of automatic systems for voice pathology detection," *Biomedical Signal Processing and Control*, vol. 1, no. 2, pp. 120–128, 2006, ISSN: 17468094. DOI: 10.1016/j.bspc.2006.06.003.

[17] S. Hindurao, L. Harad, M. Babar, and P. Kachare, "Laryngeal cancer discrimination using linear predictive features," *Proceedings of the 2017 IEEE International Conference on Communication and Signal Processing, ICCSP 2017*, vol. 2018-Janua, pp. 1786–1790, 2018. DOI: 10.1109/ICCSP.2017.8286702.

[18] A. de Cheveigné and H. Kawahara, "YIN, a fundamental frequency estimator for speech and music," *The Journal of the Acoustical Society of America*, vol. 111, no. 4, pp. 1917–1930, 2002, ISSN: 0001-4966. DOI: 10.1121/1.1458024. [Online]. Available: http://asa.scitation.org/doi/10.1121/1.1458024.

[19] C. A. Rødbro, *Yin_mgc*, 2002.

[20] R. Togneri and D. Pullella, "An overview of speaker identification: Accuracy and robustness issues," *IEEE Circuits and Systems Magazine*, vol. 11, no. 2, pp. 23–61, 2011, ISSN: 1531636X. DOI: 10.1109/MCAS.2011.941079.

[21] C. M. Bishop, *Pattern Recognition and Machine Learning*. Springer Science+Business Media, 2006.

[22] E. J. Candes, X. Li, Y. Ma, and J. Wright, "Robust Principal Component Analysis?," vol. 58, no. 3, 2009, ISSN: 0899-7667. DOI: 10.1145/1970392.1970395. arXiv: 0912.3599. [Online]. Available: http://arxiv.org/abs/0912.3599.

[23] G. E. P. Box, "A General Distribution Theory for a Class of Likelihood Criteria," *Biometrika*, vol. 36, no. 3, pp. 317–346, 1949.

[24] A. Trujillo-Ortiz and R. Hernandez-Walls, *MBoxtest: Multivariate Statistical Testing for the Homogeneity of Covariance Matrices by the Box's M. A MATLAB file.* 2002. [Online]. Available: http://www.mathworks.com/%7B%5C%%7D0A%7B%5C%%7D25%20matlabcentral/fileexchange/loadFile.do?objectId=2733%7B%5C&%7DobjectType=FILE.

[25] K. Krishnamoorthy and J. Yu, "Modified Nel and Van der Merwe test for the multivariate Behrens-Fisher problem," *Statistics and Probability Letters*, vol. 66, no. 2, pp. 161–169, 2004, ISSN: 01677152. DOI: 10.1016/j.spl.2003.10.012.

Appendices

Appendix A

Additional figures and tables

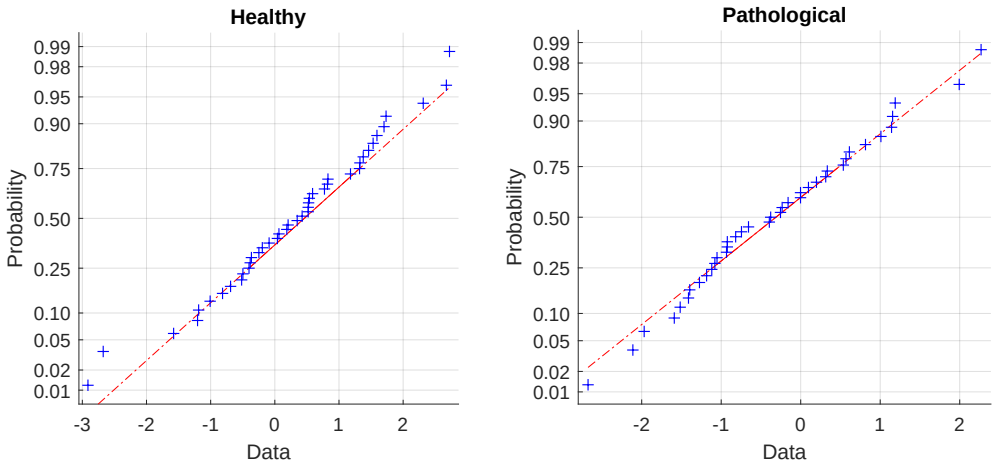


Figure A.1: Normal probability plot of the non stationary speech data projected onto the 5:th principal component.

Table A.1: Confusion matrix using classification by threshold in the space of the 5th principal component.

18	8
21	34

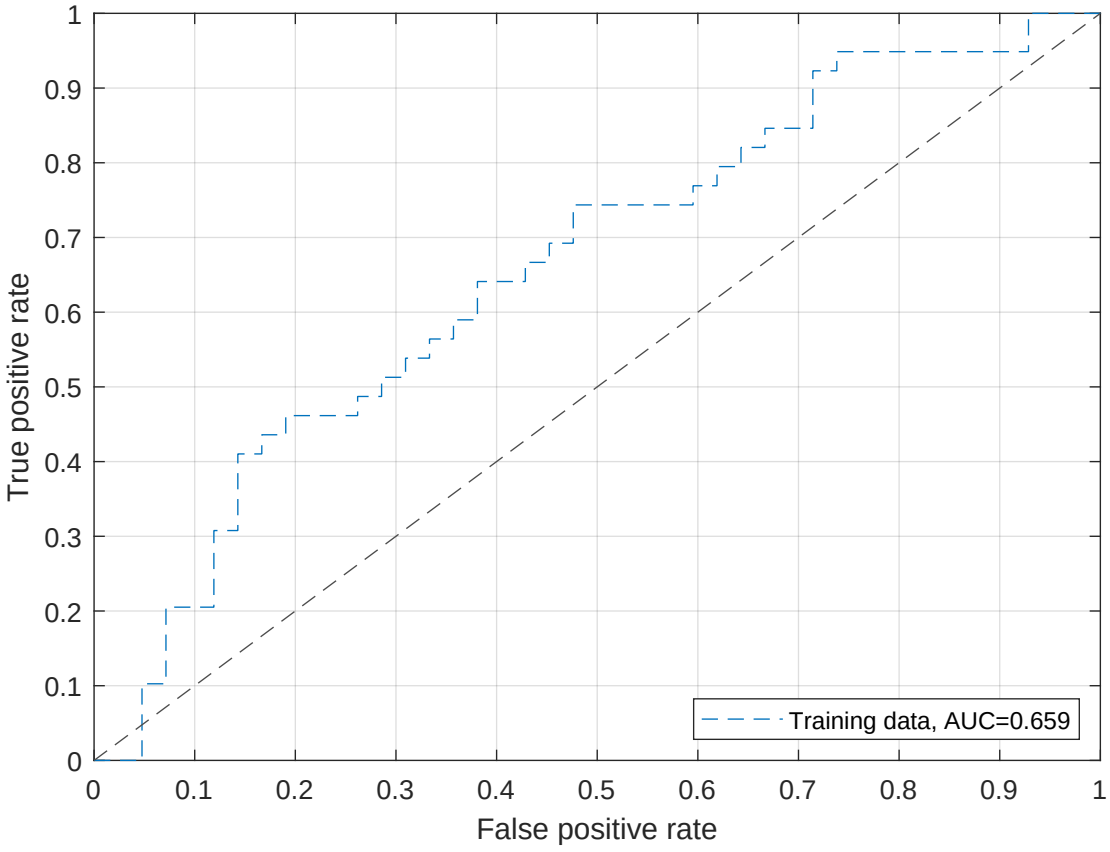


Figure A.2: ROC curve using classification by threshold on the 5th principal component.