university of groningen / campus fryslân

Automatic Detection and Severity Estimation for Oral Cancer Speech

Janay Monen

university of groningen / campus fryslân

University of Groningen

Automatic Detection and Severity Estimation for Oral Cancer Speech

Master's Thesis

To fulfill the requirements for the degree of
Master of Science in Voice Technology at
the University of Groningen under the supervision of
Dr. Vass Verkhodanova (University of Groningen)
and external supervision of
Bence M. Halpern (University of Amsterdam)

Janay Monen (s4840054)

July 9, 2022

Acknowledgments

This project would not have been possible without the great support of several people. I especially want to express my gratitude to my external supervisor, Bence Halpern, for guiding me through this whole process. Not only did he give me the opportunity to research oral cancer speech, but he also consistently provided valuable and constructive feedback that helped me get to this final point. I am truly thankful to have worked with him. The second person I want to thank is my supervisor, Dr. Vass Verkhodanova. I do not think that I would have been able to push through till the very end if it were not for her emotional support. Lastly, I want to thank my friend Christiaan for his late-night tech support sessions, and my friend Kirsten for being on this roller coaster journey with me. Because of their support I did not feel totally alone in this new but exciting challenge.

Abstract

Oral cancer (OC) surgery can cause impaired speech intelligibility by preventing the production of articulatory targets required for sounds such as plosives and alveolar sibilants (Halpern et al., 2020a; Halpern et al., 2022a; Halpern et al., 2022b). Various studies have already investigated OC speech characteristics through phonetic approaches that calculate phoneme error rates from transcription-based intelligibility assessments (Saravanan et al., 2016; Constantinescu et al., 2017), and acoustic approaches that look at formants and vowel space area (Bruijn et al., 2009; Rieger et al., 2010). Few studies, however, have looked into OC speech severity estimation (SE) and distinguishing OC speech from healthy speech through machine learning (ML) (Halpern et al., 2020a). The difference between ML and the phonetic/acoustic approaches is that ML models can automatically learn distinctions based on acoustic features and estimate quantities (e.g., severity scores), something which standard significance testing cannot achieve. Using ML for OC detection could therefore broaden our understanding of OC speech, in particular by showing us which speech features are important. Additionally, SE could assist with the tracking of speech therapy progress post-surgery (Suárez-Cunqueiro et al., 2008).

Therefore, we explored OC detection and SE through four ML models: logistic regression (LR), support vector machines (SVMs), multilayer perceptrons (MLPs) and one-dimensional convolutional neural networks (1D-CNNs). Using these models, we investigated whether (1) we can distinguish OC speech from healthy speech and (2) whether SE of OC speech based on acoustics is possible. To avoid unwanted artifacts (Halpern et al., 2020a), we collected a dataset with 6 OC patients > 1 year post-surgery and 5 healthy controls. Additionally, we gathered data for a Dutch adaptation of the Speech Handicap Index (Van den Steen et al., 2011) and used the scores as ground truth for SE. Model performances were evaluated in terms of standard accuracy, area under curve, sensitivity and specificity metrics. Our findings confirm that OC speech detection is possible with models trained on long-term average spectrum (LTAS) features. The best performance on this task was achieved with the 1D-CNN (67.41% accuracy). We also found confirmation for reliable OC speech SE, in particular for the SVM trained on Mel-frequency cepstral coefficient (MFCC) features (68.73% accuracy). These outcomes suggest that model performance may depend on factors such as task, feature type and several other factors that we address in our discussion.

Contents

List of Figures

List of Tables

1 Introduction

1.1 Motivation

Oral cancer (OC) is one of the ten most common malignant diseases that affects roughly 529,500 people every year (Rivera, 2015; Shield et al., 2017). This makes OC one of the deadliest variants of cancer. Therefore, improving the treatment of OC patients is urgent in order to increase their survival. Despite the high mortality rates, OC survivors suffer from various side effects, which include difficulties such as swallowing, speech, etc.. These side effects can have a negative and severely disabling impact on a patient's quality of life (Mathog, 1991; Epstein et al., 2001). Many of these side effects are caused by the surgical treatment (ST) of OC. This treatment involves the excision of the tumor and in some cases, crucial tissues required for oral functioning such as tongue muscles have to be removed (Korpijaakko-Huuhka, 1999). The consequences of ST may, however, be compensated for by the reconstruction of vital tissues. Nonetheless, reconstructions are not always able to fully eliminate the brought about deficits (Mathog, 1991).

The speech-related side affects resulting from ST of OC often come in the form of speech impairments that affect articulation and phonation, and are caused by the total or partial removal of the tongue. The technical term for this is partial or total glossectomy. Glossectomy can thus affect a patient's speech characteristics or speech intelligibility (SI), i.e., the degree to which the speech can be understood (Furia et al., 2001). Previous research that has investigated speech characteristics of OC patients found evidence for reduced SI (Saravanan et al., 2016). The extent of reduced SI depends on the location, size and stage of the tumor, the extent and type of surgical resection, and the method of surgical reconstruction (Pace-balzan et al., 2011). In other words, speech thus depends on the "quantity, quality and mobility of the residual oral" structures (Pace-balzan et al., 2011, p.102). Several studies have already discovered sounds that are that are commonly affected by OC treatment and thus distinct from sounds produced by healthy speakers. These include sounds such as plosives (e.g., /k/), alveolar sibilants (e.g. /s/, /z/) (Halpern et al., 2020a; Halpern et al., 2022a; Halpern et al., 2022b), palatals such as the rhotic /r/, the affricate /ʧ/ (Nicoletti et al., 2004), the fricative /ʃ/ and the alveolar lateral /l/ (Saravanan et al., 2016).

One method that has been used to assess OC speech characteristics such as described here relates to the phonetic aspects of OC speech. The phonetic aspect refers to the (human) perception of sounds, from which implications about impaired SI can be made. Examples are Saravanan et al. (2016) and Constantinescu et al. (2017), who investigated this by calculating phoneme and word error rates from transcription-based intelligibility assessments on word and sentence level. Another method to assess OC speech characteristics is through acoustic analysis, where the focus is on finding differences in acoustic measures of OC speakers and control speakers. If such differences are found, this could imply that specific sounds are affected. Examples of acoustic measures used for acoustic analysis include formants and the vowel space area (Rieger et al., 2010; van Son et al., 2018). Additionally, there is research that has used machine learning (ML) techniques. A key difference between the acoustic and phonetic approaches and ML approaches is that models based on ML can automatically learn distinctions between groups based on acoustic features. Moreover, standard significance testing on the one hand, can only tell us that there is a significant difference between groups based on a certain acoustic feature. Therefore, the choice of the acoustic feature makes the results fundamentally biased. On the other hand, in ML, a large number of features (e.g., extracted from audio samples) could be used to train a classifier that is able to determine automatically which features are important. Contrary to standard significance testing outcomes, ML methods thus generate objective findings.

Two recent examples demonstrating the success of ML methods for head and neck cancer speech are Aicha (2020) and Kim et al. (2020). They were able to distinguish speech from laryngeal cancer patients from speech of healthy controls using ML models. In contrast to this, despite the progress made by previous studies on OC speech, so far only limited research has looked into OC speech detection through ML methods. One worth mentioning, however, is Halpern et al. (2020a), who used ML methods to successfully distinguish healthy speech from OC speech. Nonetheless, much more research involving ML techniques is needed to broaden our understanding of OC speech, in particular to provide professionals with suitable and unbiased tools that can assist with the detection of (post-operative) changes in the speech of (oral) cancer patients (Kim et al., 2020). Additionally, it could help plan appropriate rehabilitative speech measures to ensure successful future interpersonal communication and well-being of OC patients Saravanan et al., 2016).

Another important aspect of OC speech pertains to the assessment thereof. When OC patients undergo speech assessments, their speech is often assigned a severity score. This notion of severity is related to the extent that speech can be understood by others and the presence of atypical voice qualities that arise from surgical OC treatment (e.g., altered F0, pitch)(Zimmermann et al., 2003). Given the social and functional impact impaired speech can bring about, severity estimation (SE) of OC speech is a crucial element of the pre- and post-treatment phase of OC. Namely, it not only enables professionals to inform patients about speech-related surgery consequences but also allows for speech monitoring (Suárez-Cunqueiro et al., 2008; Woisard et al., 2021).

There are several well-known assessment instruments that are used for OC SE such as the Speech Handicap Index (SHI) (Rinkel et al., 2008), an assessment tool that allows patients to evaluate their own speech and the impact thereof on their lives (refer to Section 2.5 for more examples). Due to the subjective element, however, such assessment tools have several disadvantages. One main issue is that they lack reliability due to varying test conditions, e.g., ecological setting, reading vs spontaneous speech. A second issue arises as a result of the different experiences among professionals with respect to speech perception (Maier et al., 2007). Additionally, bringing on a group of professionals only reduces the cost efficiency, because it takes time to assess the speech of every patient. Consequently, these issues require solutions that involve more robust evaluation methods, but there are currently only few objective tools available for OC speech SE (Woisard et al., 2021). Two examples of studies that have looked into objective OC speech SE are Windrich et al. (2008) and Woisard et al. (2021). They used a ML-based automatic speech recognition (ASR) system to estimate severity index scores of OC speech automatically, after which they compared it to scores assigned by a panel of human experts. In both cases, the automatic evaluation, which used the word recognition (WR) rate as an indication for SI, correlated closely with that of the human experts. This emphasizes effectiveness of ML and the need for more ML-based methods for the development of objective OC speech SE tools. As has been previously mentioned, a trained classifier can generate unbiased outcomes by automatically assigning importance to features derived from speech signals. Though the diagnosis of OC could never be made on the basis of ML, SE could still benefit from objective methods since the speech severity of OC patients needs to be tracked to know how speech therapy progresses, or whether speech therapy is even needed.

Based on the arguments presented here, the current research therefore focuses on OC detection and SE estimation through various ML methods. The next section mentions these methods briefly and introduces our research questions (RQs).

1.2 Research Questions

This thesis explores OC speech detection and OC speech SE from the ML perspective. The aim is to broaden the general understanding of OC and provide aid to professionals in pre- and post-surgery procedures. More specifically, we use four types of ML methods to tackle these issues:

- Logistic regression (LR);
- Support vector machine (SVM);
- Multilayer perceptron (MLP);
- Convolutional neural network (CNN).

This then leads to the first RQ that this thesis addresses:

RQ1. Is it possible to distinguish healthy speech from oral cancer speech with machine learning methods?

Based on the findings of Kim et al. (2020) (laryngeal cancer) and Halpern et al. (2020a) (OC), both of whom were were able to distinguish cancer speech from healthy speech through ML techniques, we hypothesize that it is possible to distinguish Dutch OC speech from speech of healthy Dutch controls through the above-mentioned ML methods.

Another RQ that this thesis addresses is the following:

RQ2. Is it possible to estimate severity of oral cancer speech based on acoustics with machine learning methods?

Based on Windrich et al. (2008) and Woisard et al. (2021), who successfully estimated severity index scores using ML-based ASR models, we hypothesize that it will be possible to estimate OC speech severity of through our proposed ML methods.

Lastly, besides these main RQs we also explore a sub-RQ that relates to model performance:

Sub-RQ Which machine learning method is the most suitable for determining the presence and severity of oral cancer speech in terms of performance?

Based on prior cancer research using ML methods such as Kim et al. (2020) and Halpern et al. (2020a), both of whom found an advantage of Deep Neural Networks (DNNs) over more traditional methods, we hypothesize that DNN-based models such as the CNN are most suitable for OC speech detection and SE.

1.3 Thesis Outline

The rest of the thesis is organized as follows: Chapter 2 provides the background information to get a better understanding of the topic and expands on relevant previous studies, Chapter 3 outlines our methodology, Chapter 4 presents the findings, which will be discussed in more detail in Chapter 5, together with directions for future research. We end with a conclusion in Chapter 6.

2 Background

This chapter provides an overview of information that helps to further understand the motivation behind this thesis. We start by explaining OC in Section 2.1, followed by an introduction of ML and the four methods that this thesis employs in Sections 2.2 and 2.3. Next we discuss prior research into OC speech detection and its limitations in Section 2.4, after which we continue with a similar type of discussion for OC speech SE in Section 2.5.

2.1 Oral Cancer

OC or oral squamous cell carcinoma can be defined as a malignant neoplasia, a type of abnormal and excessive tissue growth that appears on the lip or in the oral cavity (Rivera, 2015). The two principal etiological factors are generally considered the consumption of tobacco and alcohol abuse, both of which have a mutual reinforcing effect and are present in about 90% of OC cases (Dhanuthai et al., 2017; Dissanayaka et al., 2012).

To recommend suitable treatments and to better predict a patient's prognosis, establishing the stage of the OC is crucial (Rivera, 2015). This is known as staging, which is a way for professionals to determine the location of the cancer and whether it has spread to other parts of the body through a series of diagnostic tests. A common tool that professionals use to determine the cancer stage is the Tumor Node Metastasis (TNM) classification system (Appendix A), which can be used to answer the following questions (Cancernet, 2021):

- **Tumor (T):** How large is the primary tumor? Where is it located?
- **Node (N):** Has the tumor spread to the lymph nodes? If so, where and how many?
- **Metastatis (M):** Has the cancer spread to other parts of the body? If so, where and how much?

As far as treatment is concerned, the backbone of OC treatment is glossectomy, also known as surgical resection (i.e., partial removal) or removal of the entire tongue (Kademani, 2007). The oral cavity is essential for speech production, deglutition (swallowing) and mastication (chewing). Glossectomy could therefore severely compromise the oral functions and the quality of a patient's life. With respect to speech production in particular, glossectomy affects the articulatory system. To produce speech, we need to change the shape of the vocal tract and airflow (Ramoo, 2021). Our tongue plays a crucial part in this process because the position of our tongue determines which sounds we produce. For instance, the plosive /t/ is produced when our tongue touches our alveolar ridge, thereby obstruction the flow of air before releasing it with a burst. In case a part or the entire tongue is missing, producing this sound might become very difficult or impossible even (see again Halpern et al., 2020a; Halpern et al., 2022a). Similar cases arise for the production of other plosives, certain palatals, affricates and laterals that were listed in Section 1.1. Consequently, this can lead to (severe) loss of SI.

Aside from glossectomy, there are alternative common OC treatments that could affect speech aspects other than articulation. One such treatments is radiotherapy, which is a type of cancer treatment that destroys cancer cells or slows down their growth with the goal to preserve vital tissues. Previous research that has investigated the speech of (OC) patients pre- and post-surgery found that radiotherapy can cause speech impairments besides articulation issues. Namely, radiotherapy can lead to patients demonstrating dysphonia, which is often associated with hoarseness and voice quality changes, phonation issues in the form of unstable vocal fold vibrations (e.g., uncontrollable pitch), and abnormalities in acoustic measures such as jitter and shimmer (Lazarus et al., 2014). Another common treatment is chemotherapy, in which anti-cancer drugs are administered to cure cancer or reduce symptoms in an attempt to prolong life. Similar to radiotherapy, it has the goal to preserve the articulatory organs. Additionally, it is often combined with radiotherapy, which is also referred to as chemoradiotherapy. Much like the speech impairments arising from radiation therapy, chemoradiotherapy has reported similar results (for more details see Barrett et al., 2004; Mowry et al., 2006). While glossectomy thus primarily gives rise to articulation issues, radiotherapy and chemoradiotherapy can also affect voice in the form of phonation problems. Section 2.5 discusses the assessment tests commonly used to assess the affects of these treatments.

2.2 Machine Learning

Voice assistants such as Siri, Alexa and Cortana are based on ML methods that enable them to interpret and learn from human speech (Hoy, 2018). For instance, every time we use Alexa to ask about the weather or request a song to be played, Alexa learns more about our voice characteristics and speaking habits as a result of its built-in ML methods. Because of this automatic learning ability, systems based on ML can thus perform a given task automatically without there being a need for explicitly programmed instructions (Mahesh, 2018). Additionally, ML also teaches machines how to extract information from large amounts of data automatically.

ML models require robust and discriminatory features to learn distinctions between items accurately and quickly (Sharma et al., 2020). The performances of traditional ML models depend on the (type of) features in the training and test set, which is why feature extraction is a crucial part of the ML process (Sharma et al., 2020). This statement does not fully apply to end-to-end (E2E) ML approaches since these require no or only minimal feature extraction: either the raw waveform (Ravanelli and Bengio, 2018) or Mel spectrograms (e.g., Li et al., 2019) are used as input. For the purpose of this thesis, however, we will not focus on E2E approaches. If we thus assume that feature extraction is crucial for the development and final performance of a ML model, we need to understand the notion of feature extraction. Since an entire dataset is often too complex, both time-wise and computationally, they are generally not compatible with ML models. To solve this problem, feature extraction can be used to compress the original audio signal into features. This reduces the computational complexity of training while highlighting the most important characteristics of the audio. Consequently, it should result in a reliable model that can predict with high accuracy.

We have demonstrated that feature extraction is an important part of ML. However, it is important to mention that there are different ML methods for different tasks. The following section describes one such task which is also the focus of this thesis, classification.

2.3 Classification

The main goal in a classification task is to assign an input vector X to either of the K discrete classes C_k. Here, $k = 1, ..., K$. Generally speaking, the classes within a classification task are considered to be completely separate from each other. An example of this is the case of object detection for self-driving cars, in which we deal with a multi-class classification paradigm (i.e., 3 or more classes). Here, the goal is to detect different objects in a single image such as a zebra crossing, a pedestrian, road signs, etc., and put them into their corresponding class. Each object is different and can only be assigned to one single class (e.g., *pedestrian* or *road sign*), which implies that the classes are mutually exclusive. Contrary to the multi-class paradigm, if $K = 2$, the classification becomes binary. To then help determine which input belongs to which class, decision regions, of which the boundaries are referred to as decision boundaries, divide up the feature space (Bishop, 2007). To achieve this type of classification, various ML methods can be used. The next few sections describe four of these methods: LRs, SVMs, MLPs and CNNs.

2.3.1 Logistic Regression

LR is a parametric classification method (Bishop, 2007). In other words, models based on LR have a fixed number of parameters that rely on the number of input features. These input features are independent variables that predict a binary outcome, i.e., LR outputs a categorical prediction. In theory, LR is very similar to linear regression (LiR) but instead of predicting a categorical variable

(e.g., *male* or *female*), LiR predicts numeric variable (e.g., the F0 of each speaker). Additionally, rather than fitting a straight line (i.e., LiR) to our observed data, an S-shaped curve is fitted to our observed data (Figure 1). This S-shaped curve is a Sigmoid, which transforms the output into values ranging between 0 and 1 and therefore makes it easy to interpret.

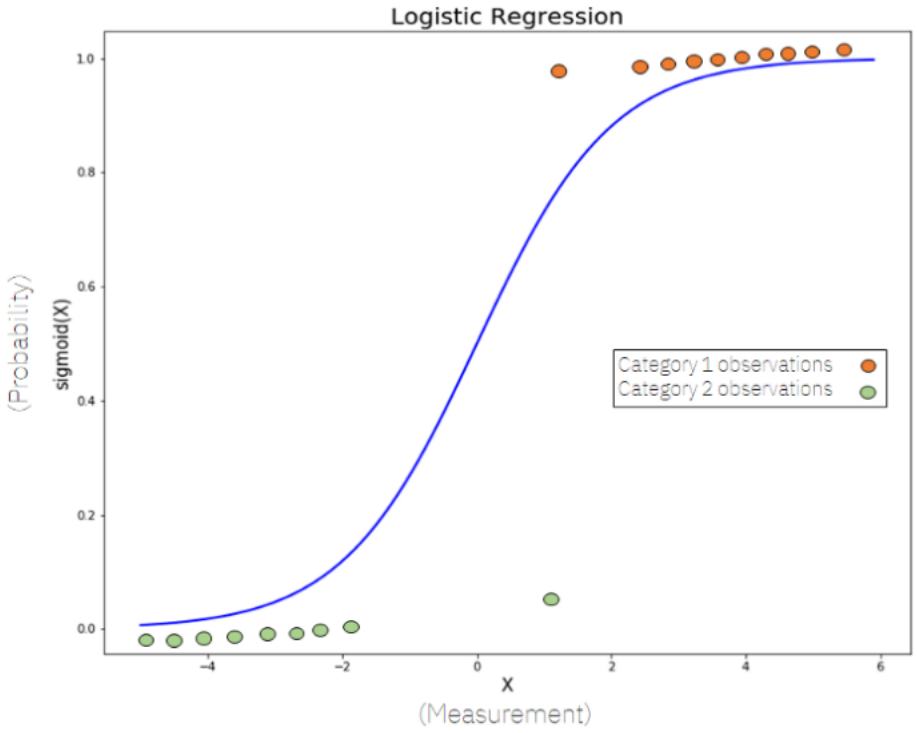

Fig. 1: *Sigmoid function represented as an S-shaped curve – reproduced from Zai (2021).*

The Sigmoid function (SF) has the following notation[1]:

$$\sigma(\eta) = \frac{1}{1+e^{-\eta}}. \tag{1}$$

As we briefly mentioned, LR outputs a binary categorical prediction. Figure 1 demonstrates that the observed data has two types of observations, with the Y-axis ranging from 0 to 1. This 0-1 range stems from LiR, but rather with an added SF that allows for the compression of observed values into [0, 1]. Moreover, the SF provides us with a value that can be interpreted as a probability, and since it is differentiable (i.e., continuous), it can be used for Gradient Descent (GD) optimization (see Phillips, 2021).

If we thus want to predict sex based on acoustic features extracted from speech signals, the first step is to calculate the weighted sum of all inputs, where Θ, z and b represent the coefficient, acoustic features and bias:

$$\sigma = \Theta \cdot z + b. \tag{2}$$

This is followed by a calculation of the probability of sex through the SF. In order to use LR to train a ML model for this classification problem, however, we need an additional step to obtain the model parameters (i.e., the weights). Therefore, we need an iterative optimization method such as

[1] e represents the Number of Euler and η represents the output. Please refer to Bishop (2007) for more details.

GD or stochastic GD (SGD) (Dokuz and Tufekci, 2021). Once the model has been trained, and the parameters have been obtained, it is then possible to make predictions about the sex of the speakers.

2.3.2 Support Vector Machine

A second ML method is the SVM, which is an easy-to-interpret method that is able to generalize well in many cases, even when there is limited data available (Hernandez et al., 2020). The idea behind SVMs for binary classification is that it seeks the most optimal decision boundary to minimize the misclassification error (Figure 2).

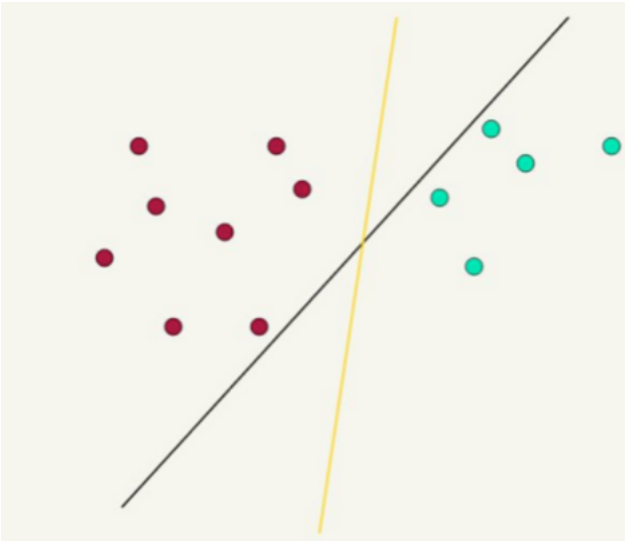

Fig. 2: *General illustration of decision boundaries for SVMs. The black and yellow lines represent examples of decision boundaries – reproduced from Zhang (2019).*

In contrast to LR, the SVM is a type of decision machine that does not output posterior probabilities (Bishop, 2007). Instead, SVMs output a set of weights w based on input features x, of which the combination can predict y. More specifically, the SVM seeks the smallest distance between the observed data and the decision boundary, which needs to be as large as possible to reduce the misclassification error (Figure 3). This is referred to as the margin or street width (see more in Bishop, 2007).

Similar to the ML approach for LR, it is important to train the SVM for the binary classification task. In this case, the SVM uses a loss function that has a regularization coefficient C. Depending on the value of C, the SVM is either hard-margin (i.e., $C = 0$) or soft-margin (large value C). In other words, the lower the value of C, the stricter the SVM is when it comes to assigning penalties to violations (i.e., misclassification) and vice versa. The loss function can then be defined as follows:

$$\lambda\|\mathbf{w}\|^2 + \left[\frac{1}{n}\sum_{i=1}^{n}\max\left(0, 1 - y_i(\mathbf{w}^T\mathbf{x}_i - b)\right)\right], \tag{3}$$

where the λ parameter (i.e., C) controls the complexity of the SVM, while w represents the errors, i.e., the distance from the decision boundary. A large enough λ will thus increase the margin size and create a line separator, whereas a smaller λ will result in a plane or hyperplane as presented in Figure 3. Refer to Dibike et al. (2001) and Burbidge and Buxton (2022) for more details.

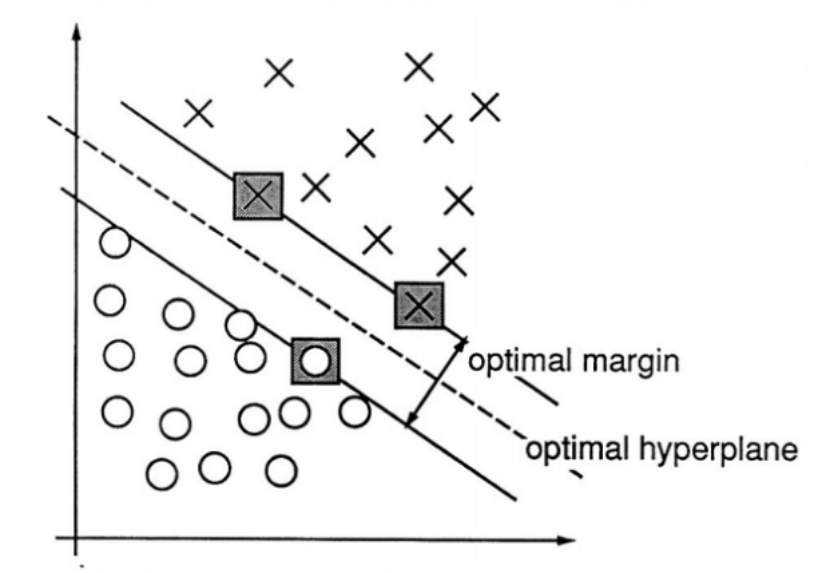

Fig. 3: *An example of a separable problem in a 2 dimensional space. The support vectors, marked with grey squares, define the margin of largest separation between the two classes – reproduced from Cortes and Vapnik (1995).*

Another important parameter is the kernel, which can be used to describe linearly non-separable data points. For instance, points inside and outside of a circle are not linearly separable and cannot be separated with a linear decision boundary. Instead, the optimal way to describe this data is through a circular decision boundary, something which only a kernel can do. The kernel function transforms the input features into the required form. Namely, it returns the inner product between two data points in a suitable feature space or window, which is the set of all possible values for a chosen set of features from the data (Amari and Wu, 2001). Furthermore, it performs a two-dimensional (2D) classification for a set of data that was originally one-dimensional (1D) (Prajapati and Patle, 2010). This increases the chance of separating them in a hyperplane. With respect to kernel type, there are several functions such as the linear or the (non-linear) polynomial kernel (Prajapati and Patle, 2010) as demonstrated in Figure 4.

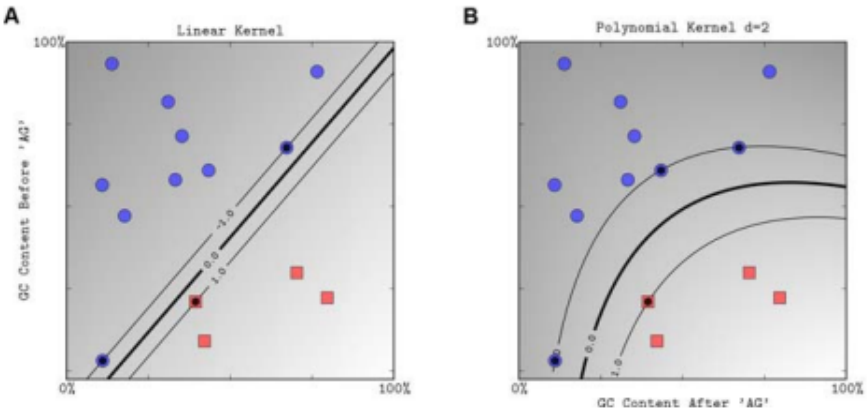

Fig. 4: *The effect of the degree of a polynomial kernel. The polynomial kernel of degree 1 leads to a linear separation (A). Higher-degree polynomial kernels allow a more flexible decision boundary (B) – reproduced from Ben-Hur et al. (2008).*

The SVM thus tries to find a balance between the number of data misclassifications and the margin of the decision boundary by using a kernel function. If we relate this back to our speech example, using a kernel function on the features results in output weights *w*, which can help us predict the sex of the speakers.

2.3.3 Multilayer Perceptron

A MLP is a type of artificial neural network (ANN) that can learn the relationship between linear and non-linear data. Additionally, is one of the most successful models in the context of pattern recognition (Bishop, 2007). The original perceptron developed by Rosenblatt is based on the functioning of neurons in the human brain (Rosenblatt, 1960). The idea is that input features are combined through a weighted sum and if that sum exceeds the predetermined threshold T, the perceptron is activated and outputs the weights to predict classes. Threshold T represents this in the form of an activation function (AF) such as Rectified Linear Unit (ReLU), Sigmoid, etc.. A downside of a single perceptron, however, is that it cannot deal with nonlinear data (Minsky and Papert, 1969). The MLP, however, consists of multiple layers of LR models that include continuous non-linearities in the hidden units, which makes it suitable for nonlinear data. The MLP also uses an AF but in the form of a feed-forward NN that performs forward propagation. Namely, each layer feeds the output of the current layer to the next layer until it reaches the output layer (Figure 5). The output is then a mean squared error.

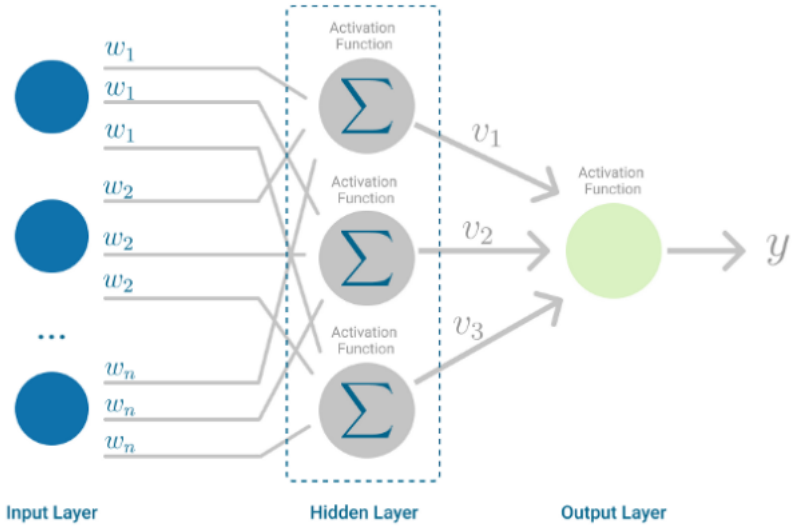

Fig. 5: *Example of a feedforward structure in a MLP – reproduced from Bento (2021).*

Moreover, backpropagation is needed for the MLP to learn weights that are required for cost minimization. Backpropagation is a type of algorithm that provides feedback about which weights are most optimal for cost minimization and computes a gradient (see Lillicrap et al., 2020).

To train a MLP-based model, we thus need an AF to compute the weights for the cost function and an optimization function such as Adaptive Moment Estimation (Adam) (Kingma and Ba, 2014) or SGD to update the cost function. Additionally, a learning rate needs to be defined to control the extent to which the weights can be adjusted for each iteration. For our binary classification example, the acoustic features are fed into the MLP. Next, the AF pushes the information through various LR layers to learn which features belong to which class (i.e., *male* or *female*), all the way until the final layer is reached. With the help of forward propagation and backpropagation, the MLP can then keep

updating the cost function to improve the model and predict the sex of the speaker.

2.3.4 Convolutional Neural Network

A CNN is a specific class of deep neural networks (DNNs) that specializes in processing data with a grid-like topology (Bishop, 2007). Examples include an array with features represented as numeric values or an image such as the one presented in Figure 6.

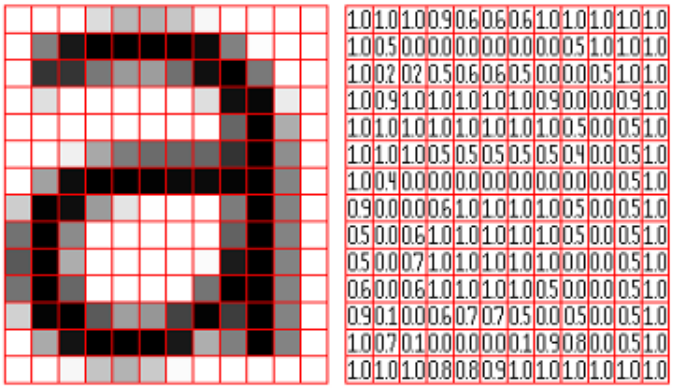

Fig. 6: *Illustration of an image as a grid of pixels represented in binary form – reproduced from Mishra (2020).*

In the human brain, every neuron has its own receptive field and together with the other neurons, they can account for the entire receptive field. Similarly, each neuron in a CNN has its own receptive field (Murugan, 2020), with the layers arranged in a way that is suitable for pattern recognition, e.g., recognizing which features characterize female and male speech.

The typical CNN consists of three distinct layers (Figure 7). The first layer type is the convolutional layer (CL), which is one of the core building blocks of a CNN. This layer performs a dot product between two matrices, with one matrix consisting of the learnable parameters (i.e., kernel) and the other matrix containing the portion of the receptive field that is restricted. Since the kernel is smaller than the original input data, the kernel needs to slide (i.e., stride) across the height and width of the input data to produce a 2D feature map[2]. The generated feature maps can be forwarded to a nonlinear AF, usually ReLU. Similar to what we described for MLPs, this AF generates output and enables a CNN to learn data that is non-linear and more complex. The second layer type in a CNN is the pooling layer. This layer slides a kernel over each channel of the feature map and compresses the output from the CL to reduce the computational complexity and number of weights needed (Albawi et al., 2017). Additionally, it ensures that the CNN can handle any variance in feature position. The final layer type of a CNN is the fully connected (FC) layer, which has the same structure as the hidden layer from the feed forward NN presented in Figure 5. The final output, i.e., the probabilities for the class predictions, can be generated through a softmax AF in the final layer of the FC layer.

With respect to training a CNN for our binary classification example, the acoustic features that were extracted from the speech signal serve as input. Consequently, through a kernel with a predefined stride and size in the CLs, the model learns important information from these features, i.e., patterns for male and female speech. Following this, the computational load and dimensionality is reduced

[2]Refer to Bishop (2007) and Albawi et al. (2017) for more details on input, output dimensions and additional model parameters

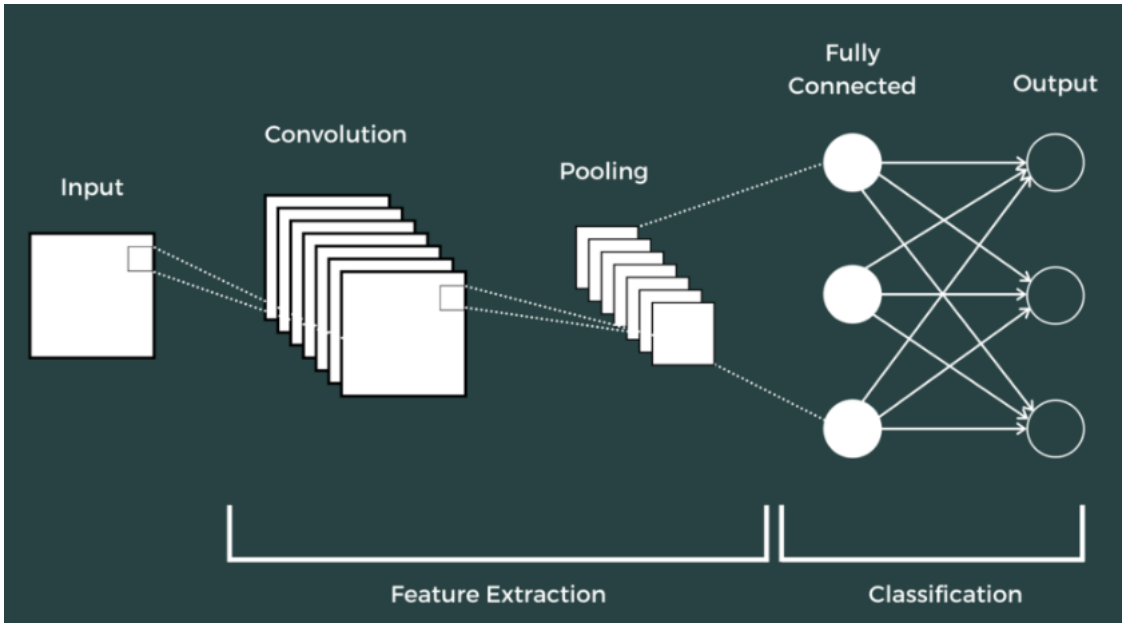

Fig. 7: *Illustration of a typical CNN structure – reproduced from Clickreader (2021).*

by the pooling layers, after which everything comes together in the FC layer. Here, a softmax AF translates the output into probabilities that represent either the *male* or *female* speaker class.

2.4 Relevant Research on OC Speech Detection

Section 1.1 briefly mentioned that research on OC detection using ML methods is limited. So far, only Halpern et al. (2020a) has investigated this in depth. More specifically, they collected 3 hours of spontaneous OC speech data from YouTube and compared that with roughly 4.5 hours of healthy speech data. As for the method, they used the preprocessing frontends of Kaldi (Povey et al., 2011) to calculate five different features: Mel-frequency cepstral coefficients (MFCCs), long-term average spectrum (LTAS), perceptual linear predictive coefficients (PLPs) and Pitch and phonetic posteriorgrams (PPG). Additionally, they decided on two backends. The first was a Gaussian Mixture Model (GMM) due to its widespread use in pathological speech detection (Dibazar et al., 2002; Bocklet et al., 2008) and the second was a least absolute shrinkage and selection operator (LASSO). The latter is a LiR method that allows for easy interpretation. Furthermore, they tested a Dilated Residual Network (ResNet), a type of DNN classifier that takes spectrograms as input (for achitecture details see Halpern et al., 2020a). Based on findings from previous spoofing detection studies (Lai et al., 2019; Halpern et al., 2020b), they expected it to be beneficial for pathological voice detection as well. Their overall findings for OC speech detection demonstrate that OC speech can reliably be distinguished from healthy speech using ML methods. The ResNet classifier in particular performed well, with an accuracy of 88.37% and a 57.82% chance-level baseline (for more details see Halpern et al., 2020a). Aside from that, they found that plosives and sibilants are crucial indicators for OC speech detection.

Despite this success, however, Halpern et al. (2020a) has several shortcomings. The first is concerned with source of the dataset. Namely, all OC speech data was collected from various YouTube videos. A disadvantage of this approach is that the use of YouTube data can bring about artifacts that can affect the final outcome. Examples of such artifacts are background or microphone noise, effects of different recording devices, etc.. A second shortcoming is related to the data inclusion procedure. To determine whether a video could be considered OC speech, Halpern et al. (2020a) analyzed the

content of each video and relied on the authors' experience with OC speech. As we have seen, experiences with regard to speech perception vary depending on the person and are thus very subjective (Maier et al., 2007). It is therefore possible that videos that should have been considered OC speech were disregarded and vice versa. A final limitation of pertains to the choice of development language. Previous studies have shown that English is often the focus language of automatic detection problems (Lavrentyeva et al., 2019; Plaza-del-Arco et al., 2021). However, OC is a worldwide concern and automatic detection of OC speech should therefore include languages other than English as well. More specifically, pronunciations of consonants such as stops, sibilants, etc., pitch and other acoustic features vary across languages (Hanley et al., 1966). In addition to this, English (language) systems typically outperform non-Enlgish systems (Korayem et al., 2016). OC speech detection that uses similar methods may therefore generate different and less favorable outcomes when the dataset consists of non-English languages. To counter all these shortcomings, it is important to consider several aspects. First, speech should be collected in a controlled setting to avoid unwanted noise and recording artifacts. If this is the case, every participant can be recorded with the same devices and noise can, for instance, be accounted for by recording in a sound-proof room. Second, OC patients should be recruited on the basis of an official OC diagnosis. This ensures that the OC speech dataset consists exclusively of OC speech and does not rely on the expertise of researchers. Third, it is important to use datasets containing non-English data to make OC speech detection methods suitable for all languages.

Another aspect to consider relates to the choice of method. Namely, it is worth exploring different ML methods for automatic OC speech detection, especially since the techniques used in Halpern et al. (2020a) are still baselines and might not be the most optimal for this task. Various research has shown that LR (Huang et al., 2016 – various speech pathologies), SVMs (Hernandez et al., 2020 – dysarthria), ANNs and 1D-CNNs (Kim et al., 2020 – laryngeal cancer) are successful in the automatic detection of pathological speech with performances well-above chance-level. For that reason, we believe that the use of these four methods could also be useful for OC speech detection.

2.5 Relevant Research on OC Speech SE

Although SE plays a crucial role in clinical practice and the methods of various research on pathological speech, there is no universally accepted definition of speech severity (Stipancic et al., 2021). In the pathological speech literature, speech severity is often determined by SI, i.e., how well speech can be understood by others (Maier et al., 2007; Windrich et al., 2008; Sussman and Tjaden, 2012; Tjaden et al., 2014). Additionally, there is a small number of studies that have used voice-based metrics such as speaking rate and duration as indicators for speech severity (Shellikeri et al., 2016; Hernandez et al., 2020). Although it is a much less prevalent proxy for severity, the use of self-reported outcomes to group the speech of pathological speakers has shown to be successful as well (Allison et al., 2017; Yunusova et al., 2016). With respect to OC speech SE, however, this type of approach is limited (Woisard et al., 2021), which is why we explore this further. Below we address such an approach, together with other common OC speech SE tools and relevant work.

Section 1.1 and 2.1 described how OC patients often suffer from speech impairments after (surgical) treatment. This in turn affects the degree with which they are able to pronounce certain sounds. There are several common assessment tests used to estimate the severity of OC speech. One tool used by clinicians and researchers worldwide is the grade, roughness, breathiness, asthenia, strain (GRBAS) scale, which assesses the voice of patients in terms of roughness, degree of dysphonia, breathiness, etc. (Nemr et al., 2012). More specifically, it considers the severity of a vocal disorder using a scale with constant intervals. The Consensus Auditory Perceptual Evaluation—Voice (CAPE-V) is

similar to the GRBAS scale but it also looks at pitch and loudness (Nemr et al., 2012). Additionally, rather than a severity scale with regular intervals, the CAPE-V scale represents the severity level (i.e., *mild*, *moderate*,*severe*) through an asymmetric scale. The SHI is also one of such assessment tools, but as we have briefly seen in Chapter 1.1, it assesses the speech rather than the voice of OC patients. The questionnaire was originally developed by Rinkel et al. (2008) to assess speech problems specific to OC and oropharyngeal cancer patients (Appendix B). The SHI consists of 30 items that are modelled based on the Voice Handicap Index (Jacobson et al., 1997). Furthermore, different from the GRBAS and CAPE-V scale outcomes, the outcomes of the SHI are all reported by the patients themselves rather than assessed by a professional. The general idea behind the SHI is that a patient's self-view of their speech might be more reflective of their quality of life than an external view (i.e., a professional). Aside from the English SHI questionnaire, Van den Steen et al. (2011) created a Dutch validated version of the original SHI to make it more accessible to non-English speakers (Appendix C and D).

Although the assessment tools described above allow for OC speech assessment, they are subjective in nature and prone to errors that result from subjective methods. This can generate biased speech severity scores that do not represent all the facts and are thus not in the best interest of the OC patients. To track the speech (progress) of OC patients pre- and/or post-surgery (Kim et al., 2020), objective speech SE tools could be useful. From Section 1.1 and 2.1, however, we know that research on this topic is limited. To the best of our knowledge, there are three studies that have attempted speech SE for OC patients through ASR-based approaches. The first is Maier et al. (2007), who investigated speech severity by looking at the SI of OC patients post-surgery. They used an ASR system based on semi-continuous Hidden Markov Models (HMMs), which is a statistical approach used to model acoustic signals (see Riedhammer et al., 2012; Maier et al., 2007). Based on a reading task, the ASR system was then able to calculate a word recognition (WR) rate that represents the patient's intelligibility score. This was compared to the intelligibility scores from a panel of experts that had been asked to do a speech assessment of the OC patients. Results from a Pearson correlation test showed a strong negative correlation ($r = -0.92$; $p < 0.01$) between the experts' rating and the automatic speech assessment. Additionally, the ASR system showed less variance when compared to the human experts, the latter of which had higher variance within their own group. Therefore, Maier et al. (2007) concluded that the performance of the ASR system was better and more reliable. This implies a superiority of objective over subjective tools. Similarly, Windrich et al. (2008) investigated the SI of OC patients post-surgery using similar experiments. Namely, they also used an ASR system based on HMMs (see Stemmer, 2005) to recognize read OC speech. Additionally, they used WR rate to calculate SI. A Pearson correlation test demonstrated a strong negative correlation ($r = –0.93$; $p < 0.01$) between the perceptually judged intelligibility scores (i.e., from a panel of experts) and those calculated by the ASR system. Based on these outcomes, they therefore emphasize the benefit and need for more objective speech SE tools. The final study that investigated SE for OC speech is Woisard et al. (2021). Contrary to Maier et al. (2007) and Windrich et al. (2008), they focused on voice characteristics (i.e., phonation) and SI to estimate severity scores. This would allow them to classify French OC patients into three categories: *mild, moderate* and *severe*. They also used an ASR system but based on LiR. The method consisted of five speech production tasks (see Woisard et al., 2021), among which were a reading task and a task where patients were asked to produce several pseudowords. These are non-existent words that follow an expected pattern depending on the chosen language. Aside from the production tasks, they also collected data from the SHI and Phonation Handicap Index, which is another self-assessment tool that is similar in structure to the SHI but assesses voice rather than speech (Fichaux-Bourin et al., 2009). The scores from these questionnaires were compared to the outcomes of the ASR system. A comparison of the automatically obtained

severity scores with the perceptual severity scores resulted in a strong positive correlation ($r = 0.87$; $p < 0.001$). Additionally, they found that ASR-based (voice) measures contributed the most to speech SE and classification, in particular the automatic average normalized likelihood scores (see Woisard et al., 2021). This again stresses the importance of ML-based speech SE assessment over subjective speech SE assessment.

Although methods for OC speech SE have thus proven to be successful, there are some shortcomings that need to be mentioned. First, the three studies described above have mainly focused on ML methods such as LiR (Woisard et al., 2021) and HMMs (Maier et al., 2007; Windrich et al., 2008). However, current state-of-the-art pathological speech SE methods also include (D)NNs that have been shown to outperform models based on traditional methods (Hernandez et al., 2020; Joshy and Rajan, 2022). Second, none of the research involving OC speech SE has addressed the effects of different ML methods on the SE such as Halpern et al. (2020a) did for OC speech detection. Therefore, we believe that it is important to test OC speech SE through a variety of traditional and state-of-the-art ML methods. Based on previous research that has been successful for pathological speech detection with LR (Huang et al., 2016 – various speech pathologies), ANNs and 1D-CNNs (Kim et al., 2020 – laryngeal cancer), and pathological speech SE with SVMs (Hernandez et al., 2020 – dysarthria), we then expect that these ML methods will also be useful for OC speech SE.

3 Method

In this chapter we describe the methods used to conduct experiments for our proposed RQs. In section 3.1 we elaborate on the methodology conducted prior to this research, which includes a description and motivation for the dataset (Sections 3.1.1 and 3.1.2), followed by a description of the data collection procedure (Section 3.1.3).

Section 3.2 presents the methods that enable us to answer the RQs from Section 1.2. To conduct our experiments, we first discuss the data preprocessing and feature extraction process (Section 3.2.1), after which we motivate our data selection approaches (Section 3.2.2). Next, we provide information on model parameters and architecture of our four chosen ML methods: LR, SVM, MLP and CNN (Section 3.2.3). Lastly, since we established that there is a need for models that can recognize and estimate OC speech (severity), it is important to evaluate overall model performances through various standardized metrics. Section 3.3 expands on this topic.

3.1 Methodology Prior to This Study

The method described in this section is part of a larger project called *Articulation and coordination of speech after treatment for oral cancer* and received ethical clearance (NL76137.042.20) (Halpern et al., 2022b). The participant recruitment and data collection procedures were completed prior to this research in collaboration with the hospital staff at the Universitair Medisch Centrum (UMC) Groningen. However, more data has been collected since then. The following sections discuss the main components of this process.

3.1.1 Dataset: Participants

Halpern et al. (2022b) collected a speech corpus with voice recordings from eleven native Dutch speakers (Table 1 and Appendix E), with ages ranging from 47 to 77. Six of the speakers, three male and three female, were previously diagnosed with OC. Additionally, each of the OC speakers was at least one year post-surgery before partaking in the research. A total of three OC patients underwent jaw surgery while the remaining group received tongue surgery to excise the tumor. Among the group that underwent tongue surgery, one patient also underwent reconstructive surgery. To allow for comparison with the oral cancer group, five healthy controls, three females and two males, were recruited as well.

Table 1: *OC and control (CON) participants as collected by Halpern et al. (2022b).*

Sex	***OC***	***CON***
Male	3	2
Female	3	3
Total	**6**	**5**

3.1.2 Dataset: Stimuli and The SHI

The stimuli used to collect speech recordings consist of a series of Dutch sentences from three sources (Appendix F). The first is the Wablieft newspaper corpus (Vandeghinste et al., 2019), which is an open-source text corpus that contains two million words from an easy-to-read Belgian newspaper, written entirely in Dutch. Halpern et al. (2022b) selected sentences so that all Dutch phonemes were included, in particular those containing plosives. A reason for this was that prior research has demonstrated that OC patients struggle to produce these sounds (Halpern et al., 2020a; Halpern et al., 2022a). The second source is a set of sentences from six Dutch texts that are often used for the assessment of speech impairments. The length of the texts differ, but they are all considered within the reading level of the speaker[3]. Lastly, since Halpern et al. (2022b) investigated the "phoneme-level manipulation capability of" an "articulatory synthesis framework", they also incorporated a set of custom sentences that contained 5 different target words in the carrier phrase, a common tool in speech therapy (Shelton and Garves, 1985). Carrier phrases are phrases where all the words of a phrase, except for one, are similar. Halpern et al. (2022b) designed the sentences in such a way that they had an identical CVC structure. Together with these custom sentences, the entire dataset contains a total of 227 sentences (Table 2).

[3]More information on the motivation for the chosen texts can be found here.

Table 2: *An overview of the total number of stimuli utterances (Utt.) per category as presented in Halpern et al. (2022b).*

Source	**Utt.**
Wablieft	76
Papa en Marloes	8
Man uit Finland	14
Noordenwind	8
Els gaat naar markt	10
Meneer van Dam	6
Jorinde en Joringel	80
Custom (repeated 5x)	25
Total	**227**

3.1.3 Data Collection

To record the speech corpus, Halpern et al. (2022b) used a Sennheiser ME66 microphone at a sampling frequency of 22,050 Hz that was set up in a sound-proof recording booth. Before starting the recording session, each participant was attached to an NDI-VOX electromagnetic articulograph. Consequently, they were asked to read all 227 stimuli out loud (Appendix F). To accommodate the participants to the sensors before the official recording session, Halpern et al. (2022b) also recorded some spontaneous speech where participants talked about their day. In addition to this, OC patients filled out a Dutch adaptation of the SHI to evaluate their speech and the impact thereof on their lives before the start of the recording session. This SHI is based on the adaptation of Van den Steen et al. (2011), but it contains an additional question related to the impact of the speech impairment on the patient's daily life. They could rate this either *none*, *slightly*, *average* or *a lot* (Figure 8). Table 3 reports the outcomes of the SHI for every participant on a scale of 0 to 60. Please note, however, that the score for PT2 is followed by a *?* as this patient left question E4 empty.

Table 3: *SHI scores of 6 OC patients collected by Halpern et al. (2022b).*

Patient	***Score***
PT1	24
PT2	17?
PT3	29
PT4	6
PT5	13
PT6	31

In total, Halpern et al. (2022b) collected around 330 minutes of read speech data from the eleven participants, with roughly 30 minutes per participant. Of those voice recordings, the majority had a duration shorter than 10 seconds. Moreover, all recordings were stored in *.wav* format and assigned with an identifier that made it easy to recognize which recording belongs to which source.

SPRAAKPROBLEMEN IN UW DAGELIJKS LEVEN

ID:_____ Score: P .../20

F .../20

E .../20

Dit zijn beweringen die veel mensen gebruikt hebben om hun spraak en de gevolgen van hun spraak op hun leven te beschrijven. Zet een kruisje bij dat antwoord dat aangeeft hoe dikwijls u dezelfde ervaring heeft.

		NOOIT	BIJNA NOOIT	SOMS	BIJNA ALTIJD	ALTIJD
P1	De snelheid waarmee ik praat is veranderd.					
P2	Ik heb het moeilijk om met mijn stem emoties uit te drukken.					
P3	Ik heb moeilijkheden om goed te articuleren wanneer ik praat.					
P4	Ik moet een inspanning doen om te praten.					
P5	Ik ben buiten adem als ik praat.					
F1	Ik heb het moeilijk om mondeling uit te drukken wat ik nodig heb (eten, drinken, WC,…).					
F2	Ik schaam me om mijn gedachten en ideeën uit te drukken.					
F3	Ik heb het moeilijk om te communiceren met mensen die ik niet goed ken.					
F4	Door mijn spraakprobleem vraagt men mij vaak om iets te herhalen.					
F5	Ik vermijd gesprekken met mijn familie, vrienden, buren.					
E1	Ik lijd onder mijn manier van praten.					
E2	Mijn spraakmoeilijkheden beperken mijn persoonlijk en sociaal leven.					
E3	Ik vind dat anderen mijn spraakproblemen niet begrijpen.					
E4	Mensen lijken geïrriteerd door mijn spraakproblemen.					
E5	Ik voel me gehandicapt omwille van mijn spraakmoeilijkheden.					

Omcirkel wat voor u van toepassing is:

Hoeveel last ondervindt u van uw spraakprobleem in het dagelijkse leven? geen – licht – matig - veel

Fig. 8: *Dutch adaptation of the Speech Handicap Index (SHI) used by Halpern et al. (2022b)*

3.2 Approach of The Current Study

The current section reports the methods we used to explore OC speech detection and OC speech SE[4].

3.2.1 Data Preprocessing and Feature Extraction

All audio files were converted from stereo to mono and downsampled to 16,000 Hz. For research purposes, we used only the newspaper and text stimuli (i.e. 202 recordings). For feature extraction, we chose to explore two types of features based on Halpern et al. (2020a). First we extracted our chosen baseline features, MFCCs, which are features that generally represent vocal-tract information.

[4]Once the source code is available it can be accessed here under the name *OC-Classification*.

The second feature we extracted is LTAS, which is a type of voice quality measurement that is used in the early stages of pathological speech detection (Master et al., 2006; Smith and Goberman, 2014), and to track the impact of surgical treatment or speech therapy on a patient's voice quality (Tanner et al., 2005). Both features were extracted with the librosa package (McFee et al., 2015a) in Python (version 3.9.12). For MFCC feature extraction, all recordings were reflection padded to the size of the longest recording in the dataset (i.e., 15 seconds). More specifically, if an audio file was shorter than 15 seconds, we calculated the remaining samples (i.e., missing seconds) and reflected the time series on both sides of the signal to create an audio file with the correct size (duration*sampling rate). Additionally, we used window length 1024, stride 512 and 20 MFCC coefficients to extract the MFCCs. The LTAS features, however, were extracted by calculating the mean and standard deviation (SD) of librosa spectrograms and concatenating these (window length 512; stride 256). After feature extraction, the last step was to store each of the feature matrices into a flattened vector to make them suitable for model training. Consequently, the MFCC extraction resulted in a set of 2D features with dimension length (DL) 9380. For the LTAS features, this resulted in a set of 2D features with DL 2050.

3.2.2 Data Selection: Training and Test Sets

We used two approaches to answer RQ1 and RQ2. For OC speech detection (RQ1), we included data from the Dutch control and OC speakers. Moreover, to ensure that there was an equal number of OC and control speakers in the training and test sets, we applied leave-two-speaker-out (LTSO). This is an approach where two speakers are left out of the training set and are instead used for the test set. Due to the uneven number of control and OC speakers, however, we could not pair up every OC patient with a control speaker. Therefore, we chose to leave out OC patient 4 due to the low self-reported severity score (6 out of 60 points). Additionally, for an accurate model validation, we combined LTSO with 5-fold cross validation (5FCV) to increase the reliability of the results. This divided the entire dataset into five subsets (Figure 9).

Fig. 9: *Illustration of five-fold cross validation. A given data set is split into five subsections where each fold is used as a testing set, a useful method to use all data where data is limited – reproduced from Kim et al. (2020).*

In our case, the test set for each fold thus consisted of a different OC-control speaker pair, with the remaining participants serving as the training set. Since the data set is small, we also looked at the effect of speaker severity on model performance. As we have seen in Table 3, the OC patients have

varying SHI scores, something which can potentially impact model performance even if we apply 5FCV. Therefore, instead of one 5FCV, we created four more 5FCV experiments that were designed in such a way that all speakers were paired up at least once as shown in Table 4.

Table 4: *Overview of healthy control-patient (HC-PT) speaker partitioning per fold for OC speech detection. The training and test set for each of the five experiments are provided.*

EXP1	***Test set***	***Training set***	***EXP2***	***Test set***	***Training set***
Fold1	HC1 PT5	HC2, HC3, HC4, HC5 PT1, PT2, PT3, PT6	*Fold1*	HC1 PT2	HC2, HC3, HC4, HC5 PT1, PT3, PT5, PT6
Fold2	HC2 PT6	HC1, HC3, HC4, HC5 PT1, PT2, PT3, PT5	*Fold2*	HC2 PT1	HC1, HC3, HC4, HC5 PT2, PT3, PT5, PT6
Fold3	HC3 PT1	HC1, HC2, HC4, HC5 PT2, PT3, PT4, PT6	*Fold3*	HC3 PT3	HC1, HC2, HC4, HC5 PT1, PT2, PT5, PT6
Fold4	HC4 PT3	HC1, HC2, HC3, HC5 PT1, PT2, PT5, PT6	*Fold4*	HC4 PT6	HC1, HC2, HC3, HC5 PT1, PT2, PT3, PT5
Fold5	HC5 PT2	HC1, HC2, HC3, HC4 PT1, PT3, PT5, PT6	*Fold5*	HC5 PT5	HC1, HC2, HC3, HC4 PT1, PT2, PT3, PT6
EXP3	***Test set***	***Training set***	***EXP4***	***Test set***	***Training set***
Fold1	HC1 PT6	HC2, HC3, HC4, HC5 PT1, PT2, PT3, PT5	*Fold1*	HC1 PT1	HC2, HC3, HC4, HC5 PT2, PT3, PT5, PT6
Fold2	HC2 PT3	HC1, HC3, HC4, HC5 PT1, PT2, PT5, PT6	*Fold2*	HC2 PT5	HC1, HC3, HC4, HC5 PT1, PT2, PT3, PT6
Fold3	HC3 PT2	HC2, HC3, HC4, HC5 PT1, PT3, PT5, PT6	*Fold3*	HC3 PT6	HC1, HC2, HC4, HC5 PT1, PT3, PT5, PT6
Fold4	HC4 PT5	HC1, HC2, HC3, HC5 PT1, PT2, PT3, PT6	*Fold4*	HC4 PT2	HC1, HC2, HC3, HC5 PT1, PT3, PT5, PT6
Fold5	HC5 PT1	HC1, HC2, HC3, HC4 PT2, PT3, PT5, PT6	*Fold5*	HC5 PT3	HC1, HC2, HC3, HC4 PT1, PT2, PT5, PT6
EXP5	***Test set***	***Training set***			
Fold1	HC1 PT3	HC2, HC3, HC4, HC5 PT1, PT2, PT5, PT6			
Fold2	HC2 PT2	HC1, HC3, HC4, HC5 PT1, PT3, PT5, PT6			
Fold3	HC3 PT5	HC1, HC2, HC4, HC5 PT1, PT2, PT3, PT6			
Fold4	HC4 PT1	HC1, HC2, HC3, HC5 PT2, PT3, PT5, PT6			
Fold5	HC5 PT6	HC1, HC2, HC3, HC4 PT1, PT2, PT3, PT5			

For instance, as Table 4 shows, HC1 is paired up with every PT at least once: PT5 in EXP1, PT2 in EXP2, PT6 in EXP3, PT1 in EXP4 and PT3 in EXP5.

Contrary to the above-mentioned approach, we only trained and tested our models on OC speech for OC speech SE (RQ2). Since we chose to base speech severity on the outcomes from the SHI (Table 3), and the control speakers did not complete the questionnaire, we excluded their speech data. Furthermore, because our focus is on classification and the SHI scores range from 0 to 31, we transformed the scores in such a way that they would be suitable for binary classification. We attempted a multiclass approach in a preliminary experiment but this resulted in very poor performance due to the lack of speakers. In addition to that, an increase in classes would have led to uneven class distributions whereas a two-class approach did not. As Table 5 illustrates, OC patients with an SHI score ranging between 0 and 20 received severity label 1 and severity label 2 if their score ranged between 21 and 31 (this was the highest score). Due to one missing SHI response (E4) from PT2, however, we chose to impute the score for this patient: all items from the E section were scored either 0 or 1, which is why we imputed +1 for the original score (17). Consequently, this allowed for a partitioning approach similar to that for RQ1. Namely, we also applied LTSO CV but for only one single experiment with 3 folds that have one level 1 and one level 2 patient in the test set (Table 6).

Table 5: *Overview of the severity labels (L) assigned to OC patients (PT) based on their SHI scores.*

Patient	***Score***	**L**
PT1	24	2
PT2	18	1
PT3	29	2
PT4	6	1
PT5	13	1
PT6	31	2

Table 6: *Overview of patient-patient (PT-PT) speaker partitioning per fold for OC speech SE. The training and test sets for one experiment are provided.*

Fold	***Test set***	***Training set***
Fold1	PT1-PT5	PT2, PT3, PT4, PT6
Fold2	PT2-PT3	PT1, PT4, PT5, PT6
Fold3	PT4-PT6	PT1, PT2, PT3, PT5

3.2.3 ML Methods

The current section presents information regarding the ML methods and the corresponding parameters that we selected. It should be noted, however, that we did not perform any grid search to optimize the hyperparameters. Due to our small dataset, creating a development (dev) set from the dataset would have removed crucial train data and could have deteriorated model performances. The other option was to tune on the test sets, but this increases the chances of the model overfitting on the test set. Therefore, we chose not to tune any parameters. Instead, we chose the parameters either randomly or

based on suggestions found in Scikit-learn (version 1.0.2) for LR, SVM and MLP (Pedregosa et al., 2011) and Keras (version 2.8.0) for CNN (Chollet et al., 2015).

As we will present in the next few sections, the model parameters may differ slightly depending on the classification task (i.e., OC speech detection or OC speech SE) but not feature type, because we expect that any differences in model performance can be attributed to the type of acoustic feature. With regards to the experimental setup for RQ1, however, it is also important to clarify that each model was run five times with identical parameters using LTSO 5FCV, once for each experiment. On the contrary, we ran the models only once for the 3FCV method (RQ2).

3.2.3.1 LR

The LR classification model was adopted from the Scikit-learn library (source), of which the documentation can be found here. For both RQ1 and RQ2, we selected max number of iterations 10000, solver liblinear and penalty l2. For the C parameter, however, we chose $C = 0.009$ for RQ1 and $C = 100$ for RQ2.

3.2.3.2 SVM

Similar to the LR classifier, we also adopted our SVM classification model from the Scikit-learn library (source). Relevant documentation can be found here. For both RQ1 and RQ2, we selected max number of iterations 10000 and kernel poly. For the C parameter, however, we chose $C = 0.009$ for RQ1 and $C = 85$ for RQ2.

3.2.3.3 MLP

The last Scikit-learn classification model we adopted was the MLP (source), of which relevant documentation can be found here. For both RQ1 and RQ2, we selected max number of iterations 10000 and learning rate adaptive. For RQ1, however, we added some additional parameters: batch size 64, $\alpha = 0.0009$ and early stopping (no improvement after 3 iterations).

3.2.3.4 CNN

Following Kim et al. (2020), we built a simple 1D-CNN with the Keras library (Chollet et al., 2015).

CNN architecture and parameters for RQ1

The 1D-CNN classification model for RQ1 consists of four Convolution1D layers with ReLU activation. The input shape for the first layer, however, depends on the feature type. For MFCC features this was (9380,1) and (2050,1) for LTAS features. Moreover, the first two and last two layers each have a different kernel size (width x height) and number of output channels: (3x3) and 32 for layer 1, (3x3) and 64 for layer 2, (2x2) and 128 for layer 3 and (2x2) and 256 for layer 4. This is followed by a MaxPooling1D layer with pool size 2 and two FC layers: the first layer has 128 units and ReLU activation, and the second layer has 10 units with softmax activation. Additionally, we applied dropout to reduce overfitting, once after the MaxPooling1D layer (0.25) and once after the first FC layer (0.5). For the loss function, we used sparse categorical crossentropy (refer here) and optimizer

Adam. Batch size was set to 64 and the number of epochs was 10. Similar to the approach for the MLP model, we also applied early stopping if there was no improvement after 3 epochs.

The 1D-CNN classification model for RQ2 is nearly identical to the model used for RQ1. However, we used a dropout rate of 0.25 after the first FC layer. Additionally, we changed the number of units for the second FC to 10 and the number of epochs to 7. Lastly, we applied early stopping if there was no improvement after 1 epoch.

3.3 Evaluation and Analysis

With the Scikit-learn toolkit (Pedregosa et al., 2011), we evaluated model performances on both classification tasks through the accuracy, area under curve (AUC), specificity and sensitivity metrics. The first metric, accuracy, refers to the accuracy of the classification in %. To gain insight into how each model performed, we calculated the mean accuracy and standard deviation (SD) for each fold. Consequently, this enabled us to compute an overall mean accuracy for each model. For task 1, which consisted of five experiments per model, this was achieved by taking the mean accuracy of 25 folds.

The second metric, AUC score, represents the diagnostic accuracy and predictive ability of the model. For both tasks, we calculated this score per fold and created receiver operating characteristic (ROC) curves for further evaluation. Aside from the ROC curves, we also calculated the overall sensitivity and specificity (%) per model. These two metrics provide an overview of the overall performance of these models and can be calculated in the following manner[5]:

$$Sensitivity = \frac{TP}{TP+FN} \tag{4}$$

$$Specificity = \frac{TN}{TN+FP}. \tag{5}$$

Lastly, we performed a Pearson correlation test for task 1 to investigate whether there is a significant negative or positive correlation between the test accuracy scores of the models and the SHI scores of OC patients in the test set. A significant correlation could potentially indicate that speaker severity is an important variable for overall model performance.

[5] **Abbreviations:** *TP* – true positive, *FN* – false negative, *TN* – true negative; *FP* – false positive.

4 Results

This chapter presents the performances of the models and corresponding experiments as described in Chapter 3. We first provide the outcomes of the OC speech detection task in Section 4.1. These include the model performances of the LTSO 5FCV approach for five experiments and results from a Pearson correlation test. Following these findings, we report the outcomes of one OC speech SE experiment in Section 4.2. All results are illustrated through standard accuracy, AUC, sensitivity and specificity metrics.

4.1 Task 1: OC Speech Detection

Table 7 reports the accuracies of the models for 5FCV OC speech detection. Table 8 presents a simplified version that contains the overall model means, AUC scores and sensitivity and specificity levels. The latter three are discussed in later sections. The results presented in Table 7 seem to indicate that the chosen features and speaker severity, i.e., the severity of the OC patient in a test fold, affect the accuracy of the models. From the model means (vertical), it becomes clear that on average, models trained on LTAS features outperformed models trained on MFCCs. More specifically, if we look at the mean accuracies for each model, we can conclude that the models with LTAS features always performed above the 50% chance-level baseline, whereas the models with MFCC features did not perform above chance-level in any instance. With respect to model performance per fold, however, Table 7 shows that the mean accuracy of a fold (horizontal) varies considerably. Namely, although the LTAS experiments demonstrate overall higher accuracies for each fold, there are several folds for which the models performed below chance-level. Similarly, contrary to the overall findings, some folds in the MFCC experiments did obtain accuracies above chance-level. In the next two sections, we expand further on these findings and the other evaluation metrics for each feature type.

4.1.1 Detection with MFCC Features

As we have briefly mentioned, Table 7 (vertical means) indicates that the MFCC experiments generally performed below chance-level. This suggests that averaged over 25 folds, no model could reliably detect OC speech and distinguish it from healthy speech. Interestingly, the SVM did outperform the other three models (49.62%) and is closely followed by the 1D-CNN (49.31%). Aside from the overall model accuracies, Table 7 also displays the mean accuracies for 25 individual folds (horizontal) and their corresponding control-patient pairs. Contrary to the overall findings, it becomes evident that for 4 of the 25 test folds, the models were on average able to assign the correct labels to speech of OC and control speakers: 1-F5 (55.99%), 2-F3 (62.81%), 3-F3 (70.79%) and 4-F5 with the highest obtained accuracy (78.65%). Additionally, this fold (4-F5) also contains the overall best classifier, i.e., the 1D-CNN (96.53%). As Chapter 3 described, we performed a Pearson correlation to potentially account for the discrepancies in accuracy among folds. Though the overall correlation in Table 9 is positive ($r(23) = 0.14$)[6], i.e., a higher severity score should result in a higher test accuracy and vice versa, it is a very weak correlation and not significant ($p = 0.49$). This suggests that the reason behind the considerable differences in test accuracy cannot be fully attributed to the severity level of the OC speaker. Instead, there may be other factors that affect the test accuracy (see Chapter 5 for more).

The other evaluation metrics and ROC curves for the OC speech detection task are demonstrated in Table 8 and Figure 10. In support of Table 7, among the models with MFCC features, the 1D-CNN (AUC=0.52) and SVM (AUC=0.50) were on average best able to differentiate between classes (Figure 10). In terms of sensitivity levels, however, the LR scored higher (64.85%). This indicates that if the model encountered speech of an OC patient in the test set, it also classified said speaker as such with better accuracy than the other models. For sensitivity, i.e., classifying a control speaker as a control, the SVM obtained the highest accuracy (87.62%).

[6] df refers to degrees of freedom.

Table 7: *Test set accuracies of the OC speech detection classifiers reported per fold (F) with two different features (MFCC and LTAS). The chance-level baseline is 50% and the number before F indicates the experiment (1 to 5). Additionally, best accuracy scores are emphasized in* ***bold*** *whereas worst accuracy scores are underlined. Patient-healthy control (PT-HC) pairs have also been provided, together with the SHI score and severity level (L) assigned to each patient.*

MFCCs	***4-F5***	***3-F3***	***2-F3***	***1-F5***	***2-F5***	***2-F2***	***4-F3***	***1-F3***	***4-F4***	***5-F5***	***3-F5***	***5-F3***	***5-F2***	***5-F1***	***2-F1***	***4-F1***	***3-F2***	***1-F4***	***5-F4***	***1-F2***	***2-F4***	***1-F1***	***3-F4***	***3-F1***	***4-F2***	***Mean model***	***SD***
LR	83.91%	71.04%	57.18%	47.52%	48.76%	46.53%	49.75%	40.59%	42.57%	39.60%	42.33%	50.00%	36.63%	36.14%	36.63%	30.20%	28.22%	23.02%	27.72%	24.30%	21.04%	1.48%	10.64%	3.96%	3.96%	36.11%	19.88
SVM	50.00%	50.00%	50.00%	50.00%	50.00%	49.01%	50.00%	47.73%	50.00%	50.00%	45.79%	50.00%	50.00%	50.00%	50.00%	50.50%	50.00%	50.00%	47.52%	50.00%	50.00%	50.00%	50.00%	50.00%	50.00%	**49.62%**	1.05
MLP	84.16%	76.73%	57.43%	50.50%	49.26%	42.82%	47.77%	34.41%	47.03%	41.83%	42.82%	29.95%	39.85%	34.65%	42.83%	31.44%	27.97%	26.49%	18.81%	25.74%	29.21%	23.02%	19.55%	3.47%	4.95%	37.31%	18.66
1D-CNN	**96.53%**	85.40%	86.63%	55.94%	50.00%	59.41%	50.00%	72.28%	48.02%	55.20%	49.50%	46.78%	42.08%	42.33%	27.97%	41.09%	45.79%	50.50%	46.04%	33.91%	28.47%	46.29%	22.28%	44.80%	5.45%	49.31%	20.05
Mean fold	**78.65%**	70.79%	62.81%	55.99%	49.51%	49.44%	49.38%	48.75%	46.91%	46.66%	45.11%	44.07%	42.14%	40.78%	39.36%	38.31%	38.00%	37.50%	35.02%	33.49%	32.18%	30.20%	25.62%	25.56%	16.09%		
SD	17.31	13.05	14.07	3.07	0.53	6.16	0.84	14.38	2.72	4.76	2.86	8.32	4.93	6.05	8.10	8.21	10.01	12.81	12.44	10.21	10.77	19.54	14.72	21.92	19.59		
Test HC	HC5	HC3	HC3	HC5	HC5	HC2	HC3	HC3	HC4	HC5	HC5	HC3	HC2	HC1	HC1	HC1	HC2	HC4	HC4	HC2	HC4	HC1	HC4	HC1	HC2		
Test PT	PT3	PT2	PT3	PT2	PT5	PT1	PT6	PT1	PT2	PT6	PT1	PT5	PT2	PT3	PT2	PT1	PT3	PT3	PT1	PT6	PT6	PT5	PT5	PT6	PT5		
LTAS	***3-F3***	***1-F3***	***2-F2***	***2-F1***	***4-F5***	***5-F2***	***3-F5***	***5-F3***	***5-F4***	***1-F4***	***1-F5***	***4-F2***	***5-F1***	***2-F5***	***4-F1***	***4-F4***	***3-F2***	***3-F4***	***2-F3***	***1-F2***	***4-F3***	***2-F4***	***5-F5***	***1-F1***	***3-F1***	***Mean model***	***SD***
LR	68.56%	78.47%	74.50%	66.93%	68.07%	57.18%	67.33%	75.99%	69.31%	65.10%	54.21%	66.58%	73.76%	72.77%	59.16%	51.24%	62.13%	64.36%	54.95%	55.45%	46.53%	44.06%	33.17%	34.16%	11.14%	59.00%	15.72
SVM	85.89%	62.62%	75.25%	80.94%	50.00%	73.76%	53.96%	47.78%	56.93%	50.00%	68.32%	49.75%	50.00%	50.00%	70.79%	53.22%	49.75%	50.00%	47.03%	50.74%	52.72%	50.99%	50.00%	50.00%	50.00%	57.22%	11.49
MLP	85.15%	87.38%	87.62%	77.72%	80.45%	73.51%	69.06%	85.15%	67.08%	69.06%	67.57%	77.97%	71.78%	67.82%	51.49%	59.90%	71.29%	56.69%	76.98%	52.72%	50.25%	46.04%	39.85%	35.64%	7.72%	64.72%	18.95
1D-CNN	86.14%	89.60%	80.20%	76.24%	**92.82%**	80.45%	90.35%	63.12%	77.97%	83.66%	74.26%	69.01%	67.08%	67.82%	67.82%	77.97%	58.17%	62.62%	42.82%	59.41%	58.17%	56.93%	59.92%	31.43%	11.39%	**67.41%**	18.76
Mean fold	**80.67%**	79.52%	79.39%	75.46%	72.84%	71.23%	70.18%	68.01%	67.82%	66.96%	66.09%	65.83%	65.66%	64.60%	62.32%	60.58%	60.34%	58.42%	55.45%	54.58%	51.92%	49.51%	45.74%	37.81%	20.06%		
SD	7.44	10.61	5.23	5.21	15.82	8.57	13.03	14.06	7.49	11.98	7.33	10.21	9.36	8.67	7.57	10.54	7.45	5.63	13.17	3.25	4.23	4.97	10.15	7.20	17.35		
Test HC	HC3	HC3	HC2	HC1	HC5	HC2	HC5	HC3	HC4	HC4	HC5	HC2	HC1	HC5	HC1	HC4	HC2	HC4	HC3	HC2	HC3	HC4	HC5	HC1	HC1		
Test PT	PT2	PT1	PT1	PT2	PT3	PT2	PT1	PT5	PT1	PT3	PT2	PT5	PT3	PT5	PT1	PT2	PT3	PT5	PT3	PT6	PT6	PT6	PT6	PT5	PT6		
Severity L	PT1-L2	PT2-L1	PT3-L2	PT5-L1	PT6-L2																						
SHI PT	24	18	29	13	31																						

Table 8: *Evaluation metrics table for the OC speech detection task. All values represent the mean of each model across 25 folds. The chance-level baseline for accuracy, sensitivity and specificity is 50% and AUC scores are on a scale of 0 to 1.*

MFCCs	***LR***	***SVM***	***MLP***	***1D-CNN***
Accuracy	36.11%	49.62%	37.31%	49.31%
Accuracy SD	19.88	1.05	18.66	20.05
AUC	0.29	0.50	0.31	0.52
Sensitivity	64.85%	11.58%	35.15%	44.06%
Specificity	52.97%	87.62%	38.12%	54.46%
LTAS	***LR***	***SVM***	***MLP***	***1D-CNN***
Accuracy	59.00%	57.22%	64.72%	67.41%
Accuracy SD	15.72	11.49	18.95	18.76
AUC	0.62	0.47	0.69	0.72
Sensitivity	64.85%	14.85%	67.82%	62.89%
Specificity	52.97%	99.01%	60.39%	71.29%

Table 9: *Reported Pearson correlation (r) and significance levels (p) for the MFCC experiments, where* $p = 0.005$ *is significant and* $df = 23$.

MFCCs	***r(23)***	***p***
Mean	0.14	0.49
LR	0.11	0.61
SVM	-0.05	0.80
MLP	0.06	0.77
1D-CNN	0.27	0.19

4.1.2 Detection with LTAS Features

Contrary to the MFCC experiments, the mean accuracies (vertical) for the LTAS experiments in Table 7 demonstrate that the models generally scored above the 50% chance-level baseline. This implies that averaged over 25 folds, all models were able to reliably detect OC speech and distinguish it from healthy speech. Out of the four models, the 1D-CNN obtained the highest accuracy (67.41%), followed closely by the MLP (64.72%).

With regard to the mean accuracies of the individual folds (horizontal), 3-F3 obtained the highest accuracy (80.67%). The overall best classifier, however, can be found in 4-F5: the 1D-CNN (92.82%). In contrast to this above chance-level performance, there are four folds that contradict the implication that models trained on LTAS features can reliably detect OC speech: 3-F1 (20.06%), 1-F1 (37.81%), 5-F5 (45.74%) and 2-F4 (49.51%). Again, we performed a Pearson correlation to see whether the test set accuracy correlates with the severity of the OC patient in the test set. Rather than than the weak positive correlation that was found for the MFCC experiments, the results from Table 10 demonstrate a negative correlation ($r(23) = -0.29$). This should imply that a lower severity score will result in

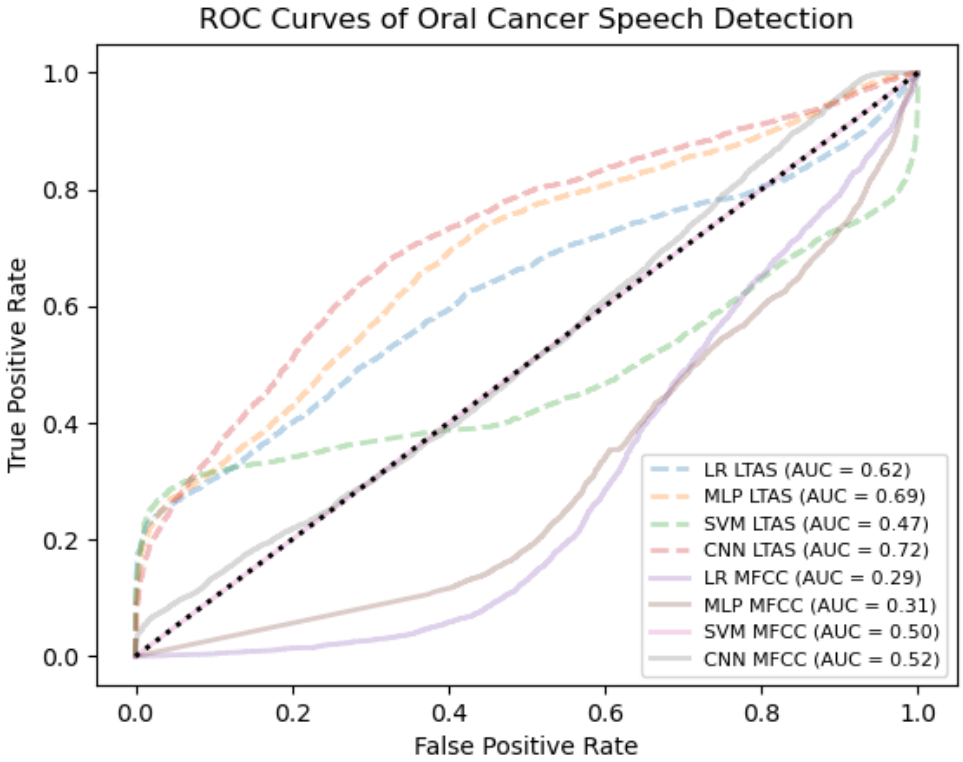

Fig. 10: *ROC curve analysis of the different models for the classification of OC speech. The positive class refers to the OC patients and the negative class refers to the healthy controls.*

a higher test accuracy and vice versa, which is quite interesting. Nonetheless, the correlation is not significant ($p = 0.16$) so we cannot conclude that the sole reason behind the discrepancies in accuracy is the severity level of the OC patient in the test set (see again Chapter 5).

Table 10: *Reported Pearson correlation (r) and significance levels (p) for the LTAS experiments, where* $p = 0.005$ *is significant and* $df = 23$.

LTAS	***r(23)***	***p***
Mean	-0.29	0.16
LR	-0.31	0.13
SVM	-0.24	0.25
MLP	-0.29	0.77
1D-CNN	-0.15	0.47

With respect to the outcomes of the other evaluation metrics in in Figure 10 and Table 8, we can conclude that the 1D-CNN (AUC=0.72) was on average best able to differentiate between classes, followed closely by the MLP (AUC=0.69) and LR (AUC=0.62). In terms of sensitivity levels, however, the MLP scored higher than the 1D-CNN (67.82%), whereas for specificity the SVM obtained the highest levels (99.01%). This suggests that the MLP on the one hand, performed best with respect to the correct classification of OC speech. The SVM on the other hand, outscored the other models in the correct classification of healthy speech.

4.2 Task 2: OC Speech SE

Table 11 reports the accuracies of the models for 3FCV OC speech SE. Table 12 presents a simplified version that contains the overall model means, AUC scores and sensitivity and specificity levels. The latter three are again discussed in later sections. The findings reported in Table 11 suggest that accuracy depends on the feature type, model and potentially the speaker severity. Contrary to findings reported for OC speech detection, it becomes clear from Table 11 that the MFCC rather than the LTAS experiment obtained higher mean accuracies (vertical). Whereas the LTAS experiments never reached above chance-level (50%) performance, the MFCC experiments did in several instances. Interestingly, however, the accuracies of the individual folds indicate that there was an instance where the classification was successful for the LTAS experiment, as well as an unsuccessful instance for the MFCC experiment. The following two sections will discuss each of these outcomes.

Table 11: *Test set accuracies of the OC speech SE classifiers reported per Fold (F) with two different features (MFCC and LTAS). The chance-level baseline is 50%. Best accuracy scores are emphasized in* ***bold*** *whereas worst accuracy scores are underlined. Patient (PT) pairs have also been provided, together with the SHI score and severity level (L) assigned to each patient.*

MFCCs	***F3***	***F1***	***F2***	***Mean model***	***SD***
LR	78.71%	50.74%	44.06%	57.84%	0.18
SVM	77.23%	**90.84%**	38.12%	**68.73%**	0.27
MLP	53.22%	50.00%	39.36%	47.52%	0.07
1D-CNN	46.04%	50.00%	43.07%	46.37%	0.03
Mean fold	**63.80%**	60.40%	41.15%		
SD	14.41	17.58	2.48		
Test PT1	PT4	PT1	PT2		
SHI/Severity PT1	6 (L1)	24 (L2)	17 (L1)		
Test PT2	PT6	PT5	PT3		
SHI/Severity PT2	31 (L2)	13 (L1)	29 (L2)		
LTAS	***F1***	***F3***	***F2***	***Mean model***	***SD***
LR	50.00%	49.26%	7.67%	35.64%	0.24
SVM	**52.97%**	42.33%	0.99%	32.10%	0.27
MLP	50.00%	50.25%	47.03%	**49.09%**	0.02
1D-CNN	50.00%	45.30%	50.00%	48.43%	0.03
Mean fold	**50.74%**	46.79%	26.22%		
SD	1.29	3.17	22.24		
Test PT1	PT1	PT4	PT2		
SHI/Severity PT1	24 (L2)	6 (L1)	17 (L1)		
Test PT2	PT5	PT6	PT3		
SHI/Severity PT2	13 (L1)	31 (L2)	29 (L2)		

Table 12: *Evaluation metrics table for the OC speech SE task. All values represent the mean of each model across three folds. The chance-level baseline for accuracy, sensitivity and specificity is 50% and AUC scores are on a scale of 0 to 1.*

MFCCs	***LR***	***SVM***	***MLP***	***1D-CNN***
Accuracy	57.84%	68.73%	47.52%	46.37%
Accuracy SD	15.00	22.35	5.92	2.84
AUC	0.64	0.68	0.54	0.60
Sensitivity	50.0.0%	71.29%	48.02%	29.70%
Specificity	65.35%	65.84%	46.53%	95.54%
LTAS	***LR***	***SVM***	***MLP***	***1D-CNN***
Accuracy	35.64%	32.10%	49.09%	48.43%
Accuracy SD	19.78	22.42	2.22	1.46
AUC	0.23	0.21	0.31	0.35
Sensitivity	5.45%	6.93%	33.17%	33.17%
Specificity	65.35%	56.93%	64.85%	66.34%

4.2.1 SE with MFCC Features

As Table 11 illustrates, the SVM (68.73%), followed by the LR (57.84%) were the only models that performed above chance-level (vertical means). This suggests that these models were overall able to assign the correct severity label (1 or 2) to the OC speech. The other two models, the MLP (47.52%) and 1D-CNN (46.37%) were not able to do this reliably. Additionally, as Chapter 3 has explained, each fold had a different patient pair in the test set for this task, one level 1 and one level 2 OC patient. Based on the fold means presented in 11, the findings indicate that F1 (60.40%) and F3 (63.80%) both obtained above chance-level accuracies, whereas fold F2 (41.15%) failed to do so. Unfortunately, due to the small sample size, we were not able to calculate a Pearson correlation to investigate the effect of severity on test accuracy. However, these findings do seem to suggest that patient pair PT4-PT6 (F3) was the most optimal pairing, followed by PT1-PT5 (F1). Patient pair PT2-PT3, however, resulted in a rather poor accuracy (41.15%). A further interesting point that arises from Table 11 pertains to the accuracy the SVM from F1. Though F3 was on average the most optimal, the SVM in F1 obtained the highest classification accuracy (90.84%). This seems to suggest that the patient pair from fold F1 in the test set with a SVM results in the most reliable classifier for OC speech SE. Contrary to this, Table 11 also seems to suggest that if we were to use the patient pair from F2 in the test set, the same model would achieve the lowest accuracy score out of all four models (i.e., 38.12%).

Table 12 and Figure 11 present the other evaluation metrics and ROC curves for the OC speech SE task. These results support the findings reported in Table 11. Namely, the SVM (AUC=0.68) and LR (AUC=0.64) were overall best able to differentiate between severity classes. Regarding the sensitivity levels, i.e., classifying a level 2 patient as level 2, the SVM obtained the best results (71.29%). However, for specificity, i.e., classifying a level 1 patient as level 1, the 1D-CNN (95.54%) outscored the other models.

4.2.2 SE with LTAS Features

Contrary to the outcomes of the MFCC experiment, not a single LTAS model could reliably estimate the severity level based on the OC speech (Table 11). Additionally, the best-performing models were the MLP (49.09%) and 1D-CNN (48.43%) rather than the LR (35.64%) and SVM (32.10%). This is the opposite of what we found for the MFCC experiment and it suggests that the type of feature input can affect the model performance considerably. Though these findings cannot be supported by a Pearson correlation test either, a closer look at the folds suggests that test set patient pair PT1-PT5 from F1 results in above-chance level accuracy (50.74%). The other pairings failed to do so: F3, PT4-PT6 (46.79%) and F2, PT2-PT3 (26.22%). The poor accuracy of F2 is in agreement with the MFCC findings. Another noteworthy finding is that not the MLP but the SVM from F1 obtained the overall highest accuracy (52.97%). This implies that the patient pair from F1 results in the most reliably LTAS classifier for OC speech SE. Interestingly, the same model, but with the F2 patient pair, also obtains the worst possible accuracy (0.99%). Both of these findings are in agreement with the findings from the MFCC experiment.

Table 12 and Figure 11, however, demonstrate that the 1D-CNN obtained the highest sensitivity levels (66.34%) and was overall best able to differentiate between OC severity levels (AUC=0.35), followed by the MLP (AUC=0.31). Additionally, the 1D-CNN and MLP also had the best specificity levels (33.17%). These findings suggest that the 1D-CNN was better at identifying level 2 OC patients as level 2, whereas both the 1D-CNN and MLP models outscored other models in terms of identifying level 1 patients as level 1.

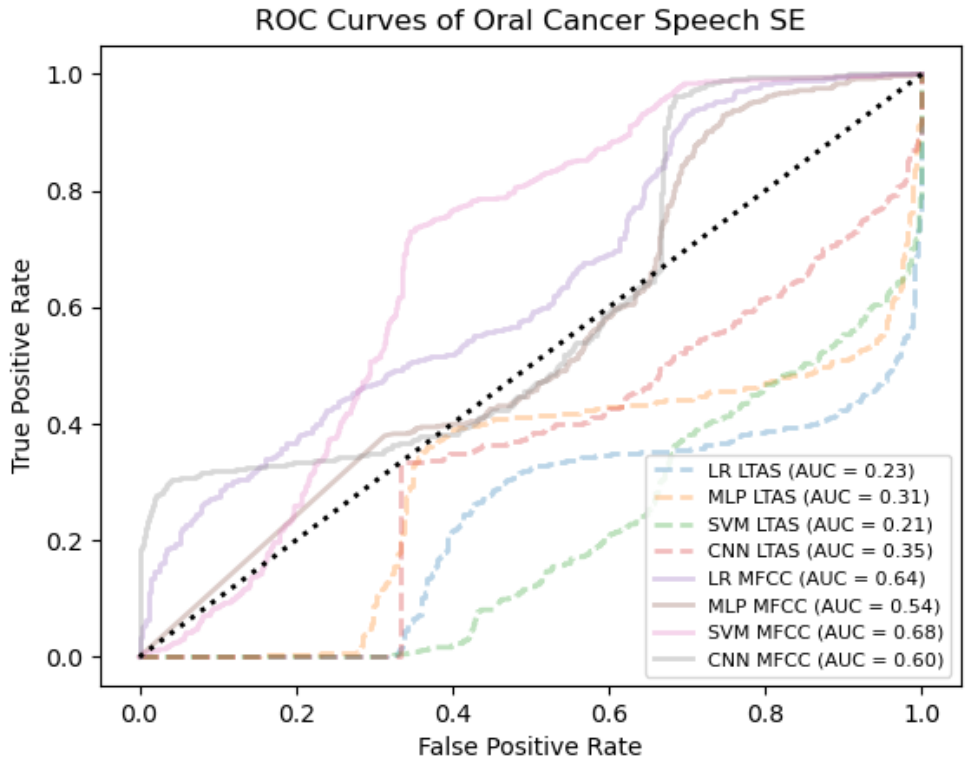

Fig. 11: *ROC curve analysis of the different models for the SE of OC speech. The positive class refers to the level 2 patients and the negative class refers to the level 1 patients.*

5 Discussion

The current chapter discusses the results reported in Chapter 4 in detail. We start by answering and discussing our first RQ: *Is it possible to distinguish healthy speech from oral cancer speech with machine learning models?* in Section 5.1. This is followed by Section 5.2, which provides an answer to our second RQ: *Is it possible to estimate severity of oral cancer speech based on acoustics with machine learning models?*. Finally, we end with a discussion of our limitations and suggestions for future research in Section 5.3.

5.1 OC Speech Detection

5.1.1 *Is OC Speech Detection Possible Using ML Methods?*

One of our main aims was to assess whether it is possible for different ML methods to successfully distinguish healthy speech from OC speech. In concurrence with previous research that has investigated OC speech with computational models (Halpern et al., 2020a), our findings suggest that it is possible to detect OC speech using various ML methods. However, these findings only extend to the experiments with LTAS feature input. Among the models used for this task, the LTAS 1D-CNN obtained the best mean accuracy across five experiments (67.41%). Additionally, our results are in line with previous studies that have demonstrated an advantage of DNNs (e.g., CNNs) and ANNs (e.g., MLPs) on certain datasets over SVMs for pathological speech detection (Godino-Llorente and Gomez-Vilda, 2004; Chuang et al., 2018). Although a 1D-CNN is more difficult to optimize, it has a high resolution and can deal with complex non-linear data (Kim et al., 2020). Additionally, it can understand spatial relations because of the tensor input. This allows for a higher learning capacity than models based on LR and SVMs (Akkaya and Çolakoğlu, 2019). Furthermore, MLPs have also shown high learning capacities, but in contrast to the 1D-CNN, they take vector input which limits the spatial awareness (Godino-Llorente and Gomez-Vilda, 2004). Since we used sequential data, i.e., acoustic (LTAS) features, this could therefore potentially explain why the 1D-CNN also outperformed the MLP. Nonetheless, our MLP achieved the highest sensitivity level (67.85%). To create assistive screening tools that can detect changes in (OC) speech, high sensitivity levels are a must (Kim et al., 2020). For this reason, MLPs may also be useful for OC speech detection. However, both the MLP and 1D-CNN models are more difficult to interpret than LR and SVM models. Therefore, it is necessary to use more explainable ML approaches such as layer-wise relevance propagation (Montavon et al., 2019) and Gradient-weighted Class Activation Mapping (Grad-CAM) (Choi et al., 2020) to visualize and understand the contributions of each element in the ML methods we used.

Nonetheless, to the best of our knowledge, this is one of the first studies that has attempted to detect OC speech with CNNs. Previous research on OC speech has already looked into various methods such as DNNs and LASSO regression in Halpern et al. (2020a). Furthermore, although there are works that have successfully adopted CNN approaches for voice pathology detection such as Kim et al. (2020) (laryngeal cancer) and Chuang et al. (2018) (dysphonia and reflux laryngitis), no research has yet bridged the gap between CNNs and OC speech detection. With support from our findings, we therefore believe that the use of 1D-CNNs for future OC speech detection devices should be considered.

5.1.2 The Effect of Speaker Severity on OC Speech Detection

An interesting finding is that the accuracies from the individual folds of each feature type vary considerably, regardless of mean (model) accuracy (Table 7). We have seen in Table 9 and 10 that these discrepancies in test accuracy do not significantly correlate with the severity of the OC speaker in the test set. Nonetheless, the two features that we used do present contrasting findings: the MFCC experiments report nearly exclusively (weak) positive correlations and the LTAS experiments report only (weak) negative correlations. This might still reveal something about the possibility that test set speaker severity is indeed a factor in model performance. Namely, especially the folds that contained either PT2 or PT3 in the test set obtained the best accuracy, regardless of feature type. Conversely, the presence of PT5 or PT6 in the test set generally resulted in poor performance. One cause of this may be attributed to speech severity. The general trend shows that low severity OC speech (PT5) is more difficult to detect than high severity OC speech (PT2 and PT3), with the exception of PT6 (highest

severity score). Whereas we would expect the best performance to result from a test set with PT6, our results demonstrate the exact opposite. For some folds, the test sets with PT6 performed even worse than those containing PT5. Listening to the original recordings, however, supports the given SHI scores: PT5 has good SI whereas the SI of PT6 is severely affected. It may thus be very well possible that rather than the OC speech severity, the healthiness of the control speech could have caused the discrepancies. Control speakers did not have to undergo a speech assessment, so the models may have identified features in their speech signal as features typical of OC speech. In addition to this, we believe that the small size of our dataset has could have played a large role as well. In particular because there was only data of six patients, there was not enough variety in terms of severity to train all the models accurately and account for all possible types of OC speech. A larger dataset could therefore improve the overall classification accuracy.

5.1.3 Comparison with Halpern et al. (2020a)

For a comparison with Halpern et al. (2020a), we need to focus on their LASSO and our LR models. The overall findings indicate that their LASSO-MFCC (80.88%) and LASSO-LTAS (87.37%) outperformed our LR models (MFCC: 36.11%; LTAS: 59.00%). Since there are many reasons that could explain these discrepancies, we discuss what we believe are the most important reasons below. A first reason is that our dataset consisted of self-collected speech data whereas Halpern et al. (2020a) collected an OC speech corpus from YouTube. Consequently, our sources were limited, so we collected data from whichever patient was available regardless of the severity. Conversely, because Halpern et al. (2020a) collected their data from the Internet, they were able to select whichever type of OC speech they wanted. Though the severity of the patients is not presented, this could have allowed them to include a wider variety of speaker severity and could have made the models more sensitive to the different severity levels. The second reason relates to language of choice. Namely, Halpern et al. (2020a) used English data whereas we used Dutch data. According to Hanley et al. (1966), variations in pronunciations across languages may generate different acoustic features. Consequently, the extracted acoustic features that resulted from Halpern et al. (2020a) and our study could in fact be language-dependent and focus on different aspects of OC speech. Therefore, future research should look into the effect of language on model performance, e.g., a comparison of English and Dutch OC speech. A third and more general reason is related to the model choice. Rather than using the same models as Halpern et al. (2020a), we chose to explore ML methods such as SVMs and CNNs that have been successfully applied to other pathological speech (Hernandez et al., 2020; Kim et al., 2020) but not yet to OC speech. This could then suggest that our models (minus the 1D-CNN) are less suitable for OC detection than methods presented in Halpern et al. (2020a). However, we believe that it is more likely that the discrepancies were caused by difference in the dataset (i.e., variety) and data preprocessing. This brings us to the final reason that we will address, which is related to feature extraction.

5.1.4 Poor Performance of The MFCC Features

Contrary to previous findings (Halpern et al., 2020a), our MFCC experiments generally resulted in below chance-level performance whereas LTAS experiments obtained accuracies that were significantly higher. Moreover, our results seem to suggest that MFCCs are not appropriate for OC speech detection. Kitzing (1986) demonstrates that LTAS features are better for the detection of pathological speech as they assess voice quality, whereas MFCC assess vocal tract information. Based on this argument we could then conclude that LTAS features are better for OC speech detection. However,

there are several other reasons that could potentially explain the discrepancies between the two feature types.

The first explanation concerns the feature dimensionality. Namely, since the DL of our MFCCs was relatively high (i.e., 9380) compared to the DL of the LTAS features (i.e., 2050), the poor performance could have been caused by the inability of our models to handle the high DL of the MFCCs. DL is an important parameter and should suit the (classification) problem for a model to perform well (Faris et al., 2020). We therefore suggest future research to explore the effect of DL of MFCCs on model performance. This brings us to a second possible explanation, the feature extraction step. In our case, the duration parameter determined the DL of our features. Whereas we chose a duration of 15 seconds, Halpern et al. (2020a) used 5-second speech chunks and managed to obtain good accuracies with MFCC and LTAS features. Consequently, this difference in duration could also have caused the discrepancies between their and our results. For the purpose of comparison with the LTAS features, however, we chose not to test other durations for our MFCC experiments. Another important step during feature extraction was the padding function. To obtain the final MFCCs, all voice samples went through a padding function (see Section 3.2.1). As a result, this excessive padding generated large amounts of silence for the model input. This should not be a problem for LTAS features as they are non-variable in length, but it could pose a problem for the MFCCs – they do vary in length. Another point worth mentioning is that Halpern et al. (2020a) used the Kaldi frontend (Povey et al., 2011) to calculate their features. Contrary to this, we used the Librosa package (McFee et al., 2015b) for feature extraction, something which could have caused discrepancies in terms of feature quality and consequently, final model performance.

Nonetheless, it remains unclear what the exact reason behind the poor performance of the MFCCs is. Therefore, we suggest that in addition to exploring DLs, durations and feature extraction libraries/frontends, future research regarding MFCCs should also investigate different padding techniques with our OC dataset. This is especially important since since the wrong padding function may deteriorate model performance (Qian et al., 2016).

5.2 OC Speech SE

5.2.1 *Is OC Speech SE Possible using ML Methods?*

Our second main aim was to assess whether different ML methods can successfully estimate the severity of OC speech. The findings are in accordance with Maier et al. (2007), Windrich et al. (2008) and Woisard et al. (2021), all of whom used ML-based methods in the form of ASR models to estimate severity scores for OC speech. Namely, our results suggest that it is possible to perform reliable speech SE for OC speech. Contrary to the findings for OC speech detection, however, these results only extend to the LR and SVM models from the MFCC experiment. Among these two models, the SVM demonstrated the best performance (68.73%). This implies that the use of these easy-to-interpret methods, as opposed to the more advanced methods such as 1D-CNN and MLP, are the most suitable for objective OC speech tracking. Moreover, this pertains in particular to the SVM as this model obtained the highest sensitivity levels (71.29%). Additionally, these findings also indicate that while LTAS features seem to be more crucial for OC speech detection due to their focus on voice qualities (Kitzing, 1986), MFCCs seem more relevant for OC speech SE. This implies that SE is more concerned with speech characteristics related to vocal tract information.

With respect to the model performances of the MFCC experiment, it is interesting to note that only the LR and SVM models performed above chance-level for OC speech SE. This contrasts with the superiority of the 1D-CNN in the OC speech detection task. The size of the dataset could poten-

tially explain this discrepancy. Namely, our current OC speech dataset was relatively small, i.e., 330 minutes of recorded speech. Previous research involving pathological speech (SE) has demonstrated that LR (Xue et al., 2021) and SVMs (Orozco et al., 2016; Hernandez et al., 2020; Tripathi et al., 2020) in particular consistently perform well, even if the dataset is small. So far, all previous research regarding OC speech SE used data of at least 35 OC patients (Maier et al., 2007; Windrich et al., 2008; Woisard et al., 2021), a number that is quite large compared to our six OC patients. In contrast to less advanced ML methods, (D)NNs generally require a larger amount of training data to perform well consistently (Iannizzotto et al., 2021). As we have seen for OC speech detection, the lack of a large dataset still allowed the 1D-CNN to outperform the other models. However, for OC speech SE a larger dataset might be necessary for a more robust performance.

5.2.2 The Effect of Speaker Severity on OC Speech SE

Similar to the findings that resulted from the OC speech detection task, the accuracies from the individual folds of each feature type vary considerably (Table 11). As we did not perform a Pearson correlation test, we cannot conclude whether the variety of severity levels in the test set of a fold significantly affected the accuracy. Nonetheless, our findings do seem to suggest a pattern that points in this direction. A first indication is that regardless of feature type, F2 with patient pair PT2-PT3 obtained the lowest accuracy. Based on reasons that were discussed in Section 3.2.2, we imputed the SHI score for PT2 and assigned them a severity level of 1. However, if we listen to this patient's original recordings, their speech is considerably impaired compared to that of PT4 and PT5. Moreover, it is perhaps closer to that of a level 2 speaker than a level 1 speaker. Therefore, it is very possible that the obtained accuracies for F2 were rather poor because of this methodological decision. This suggests that using self-assessed SHI scores to determine speech severity might thus not be the most ideal predictor for OC speech SE, especially since this score might not even represent the patient's actual speech severity. Contrary to the poor accuracies resulting from F2, we can conclude that F3 with patient pair PT4-PT6 obtained the overall highest accuracy in the MFCC experiment (63.80%). This is a result that we would expect since we paired a very low level 1 with a very high level 2 severity. Nonetheless, it is interesting since for OC speech detection, test folds containing patient PT6 generally did not perform well. However, taking a closer look at the type of information that the different features types represent could potentially explain these discrepancies. Namely, if we assume that MFCCs are better than LTAS features for OC speech SE, this implies that information about a person's vocal tract activity is more important than voice quality information. It is therefore possible that speech samples from PT6 contain vocal tract information that was not relevant for OC speech detection but is indeed relevant for OC speech SE.

On the contrary, these findings do not extend to the experiment with LTAS features because here, patient pair PT4-PT6 did not perform above-chance level. If we assume that LTAS features are not suitable for OC speech SE, we can expect such results. Nonetheless, there was one test fold that scored slightly above chance-level: F1 with patient pair PT1-PT5 (50.74%). This implies that OC speech SE could be possible with LTAS features. As of now, however, the reason behind this finding is unclear. Future studies will therefore have to explore which OC speech characteristics are crucial for SE.

5.2.3 Poor Performance of The MFCC and LTAS Features

In contrast to what was discussed in Section 5.1.4, the findings for OC speech SE show quite the opposite effect. More specifically, the performance of all models in the LTAS experiment was below

chance-level. Additionally, the MFCC experiment outperformed the LTAS experiment, with two models (i.e., the LR and SVM) even demonstrating above chance-level accuracies. As we mentioned in Section 5.1.4, we can attribute the poor performance of the MFCCs to duration, DL or the choice of padding function. While this can account for accuracy resulting from the MLP and the 1D-CNN, it does not explain why the LR and SVM obtained accuracies above 50%. Therefore, we believe that a possible reason could be the lack of data. Namely, we emphasized that (D)NNs generally require larger amounts of data to perform well consistently as opposed to models based on LR and SVM methods. If we make this assumption, it is then reasonable to assume that only the LR and SVM performed well. We can attribute this same explanation to the poor performance of the LTAS features as well. However, models trained on LTAS features all performed below chance-level. Nonetheless, the SVM in F1 demonstrated that above chance-level accuracy can be obtained (Table 11). If we then build on the assumption that LTAS features are less important for OC speech SE, it is possible all LTAS models, rather than just the 1D-CNN and the MLP, require more training data for OC speech SE. Other reasons behind the poor LTAS performance may also be due to the duration or DL. Our MFCCs had a high DL (i.e., 9380) compared to the LTAS features (i.e., 2050). In the case of OC speech SE, the models might thus perform better if the features have a higher DL. Based on these possible explanations, we therefore encourage future research to explore the use of MFCCs and LTAS features for OC speech SE, in particular with a larger dataset.

5.3 Limitations and Future Research

It is important to mention several limitations of this research, with the main limitation being the lack of sufficient data. In particular, the OC speech SE task could have benefited from more data. Since speaker severity seems to be an important factor, an in increase speakers that have a wider variety of severity levels, i.e., ranging from low SHI scores (e.g., 0-20) to extremely high SHI scores (e.g., 40-60), could perhaps improve overall model performance. Furthermore, since the speech of the control speakers had not been assessed beforehand, their speech could have been affected in some way that was not accounted for in our study. For that reason, future research should take both of these matters into account when selecting a dataset for OC speech SE. Another limitation related to our small dataset is that it prevented us from tuning model hyperparameters efficiently. Tuning on the test set was not ideal since the models ended up overfitting on the test set. Additionally, the creation of a development (dev) set would have forced us to remove crucial train data from the already small train set, something which would have deteriorated overall model performances. Future research should therefore focus on data collection and implement grid searches to optimize the four ML methods for the OC speech detection and speech SE tasks. A third limitation pertains to the feature extraction techniques used to conduct the experiments. As has been illustrated in Chapter 4 and discussed in Sections 5.1.4 and 5.2.3, MFCCs and LTAS features showed considerable differences depending on the task. Though we would have expected these features to perform well based on prior research (e.g., Halpern et al., 2020a; Hernandez et al., 2020; Kim et al., 2020), our outcomes contradict previous findings. Therefore, future studies that wish to expand on our research could look into more sophisticated feature extraction techniques (i.e., padding, duration, DL, toolkits) to improve model performances. Additionally, implementing feature ranking techniques should be explored to determine which feature types are most optimal for OC speech detection and OC speech SE.

Aside from these suggestions, there are several other directions that we did not explore. One is related to the sex of the speakers. In particular for OC speech detection, we had an uneven distribution of male-female speakers in both the OC (3 male, 2 female) and control (2 male, 3 female) set. Kim et al. (2020) and Fang et al. (2018) point out that the exclusion of female data from the test and

training set can significantly affect model performance. Female speakers generally have a cepstral domain with broader distributions than male speakers (Fraile et al., 2009). In some of our folds for the OC speech detection task, we paired up a female OC patient with a male control. A possibility is that for these folds, the models actually focused on the male vs female aspect of the features rather than the OC speech features. However, it is unlikely that this is the main factor that affects our model performances as the folds that only consisted of only male-male and female-female pairs performed in a similar manner (Table 7). For the OC speech SE task however, only F2 consisted of a male-male pair whereas the other folds consisted of male-female pairs, which could explain why F2 performed consistently poor (Table 11). Other than the speaker severity, the effect of sex could therefore have had an effect on the SE task. Future studies should thus be mindful of these distributions. A last recommendation from our side is to explore the effect of type of ST on model accuracy. Due to our small sample, we were not able to detect any obvious patterns that demonstrate higher accuracies for tongue, jaw surgery and/or resection. However, if there is indeed such a pattern, this may become clear with the use of a larger dataset.

6 Conclusion

The current research investigated automatic detection and speech SE of Dutch OC speech. The findings suggest that it is indeed possible to detect OC speech and thus distinguish healthy speakers from OC speakers using ML methods. The model that performed best on this task was the 1D-CNN trained on LTAS features (67.41%). Models trained on MFCC features, however, generally failed to perform above chance-level, which suggests that LTAS features are more important for OC speech detection. With regards to OC speech SE, our results demonstrate that it is possible to estimate the speech severity of OC patients with ML methods. The best score was obtained by the SVM model trained on MFCC features (68.73%). Contrary to these findings, we did not find any confirmation that it is possible to reliably estimate severity scores for the models from the LTAS experiment or the MLP and 1D-CNN models from the MFCC experiment. This implies that these methods did not succeed in assigning the correct severity labels to OC speech.

Though many questions remain with regards to which factors affect final model performance, it has become evident that factors such as feature type and the (lack of) variety in severity levels can play a role. This study has thereby introduced an alternative approach to automatic speech detection for OC in an attempt to provide better insights into OC speech characteristics post-surgery. Additionally, we presented objective speech SE techniques that could potentially be used to monitor OC speech and develop further speech treatment plans. Consequently, this newly presented evidence should encourage future research to expand on the growing body of studies on OC speech.

Bibliography

Aicha, A. B. (2020). Conventional machine learning techniques with features engineering for preventive larynx cancer detection. *2020 5th International Conference on Advanced Technologies for Signal and Image Processing (ATSIP)*, 1–5. https://doi.org/10.1109/ATSIP49331.2020.9231797

Akkaya, B., & Çolakoğlu, N. (2019). Comparison of multi-class classification algorithms on early diagnosis of heart diseases.

Albawi, S., Mohammed, T. A., & Al-Zawi, S. (2017). Understanding of a convolutional neural network. *2017 International Conference on Engineering and Technology (ICET)*, 1–6. https://doi.org/10.1109/ICEngTechnol.2017.8308186

Allison, K., Yunusova, Y., Campbell, T., Wang, J., Berry, J., & Green, J. (2017). The diagnostic utility of patient-report and speech-language pathologists' ratings for detecting the early onset of bulbar symptoms due to als. *Amyotrophic Lateral Sclerosis and Frontotemporal Degeneration*, *epub ahead of print*. https://doi.org/10.1080/21678421.2017.1303515

Amari, S.-i., & Wu, S. (2001). Improving support vector machine classifiers by modifying kernel functions. *Neural Networks*, *12*, 783–789. https://doi.org/10.1016/S0893-6080(99)00032-5

Barrett, W. L., Gluckman, J. L., Wilson, K. M., & Gleich, L. L. (2004). A comparison of treatments of squamous cell carcinoma of the base of tongue: Surgical resection combined with external radiation therapy, external radiation therapy alone, and external radiation therapy combined with interstitial radiation. *Brachytherapy*, *3*(4), 240–245. https://doi.org/https://doi.org/10.1016/j.brachy.2004.09.002

Ben-Hur, A., Ong, C. S., Sonnenburg, S., Schölkopf, B., & Rätsch, G. (2008). Support vector machines and kernels for computational biology. *PLoS computational biology*, *4*, e1000173. https://doi.org/10.1371/journal.pcbi.1000173

Bento, C. (2021). Multilayer perceptron explained with a real-life example and python code: Sentiment analysis. https://towardsdatascience.com/multilayer-perceptron-explained-with-a-real-life-example-and-python-code-sentiment-analysis-cb408ee93141

Bishop, C. M. (2007). *Pattern recognition and machine learning*. Springer.

Bocklet, T., Maier, A., Bauer, J., Burkhardt, F., & Noeth, E. (2008). Age and gender recognition for telephone applications based on gmm supervectors and support vector machines, 1605–1608. https://doi.org/10.1109/ICASSP.2008.4517932

Bruijn, M., Bosch, L., Kuik, D., Quené, H., Langendijk, J., Leemans, C., & Leeuw, I. (2009). Objective acoustic-phonetic speech analysis in patients treated for oral or oropharyngeal cancer. *Folia phoniatrica et logopaedica : official organ of the International Association of Logopedics and Phoniatrics (IALP)*, *61*, 180–7. https://doi.org/10.1159/000219953

Burbidge, R., & Buxton, B. (2022). An introduction to support vector machines for data mining, 3–15.

Cancernet. (2021). Oral and oropharyngeal cancer - stages and grades. https://www.cancer.net/cancer-types/oral-and-oropharyngeal-cancer/stages-and-grades

Choi, J., Choi, J., & Rhee, W. (2020). Interpreting neural ranking models using grad-cam. *CoRR*, *abs/2005.05768*. https://arxiv.org/abs/2005.05768

Chollet, F. et al. (2015). *Keras*. https://github.com/fchollet/keras

Chuang, Z.-Y., Yu, X.-T., Chen, J.-Y., Hsu, Y.-T., Xu, Z.-Z., Wang, C.-T., Lin, F.-C., & Fang, S.-H. (2018). Dnn-based approach to detect and classify pathological voice. *2018 IEEE International Conference on Big Data (Big Data)*, 5238–5241. https://doi.org/10.1109/BigData.2018.8622317

Clickreader. (2021). Building a convolutional neural network. https://www.theclickreader.com/building-a-convolutional-neural-network/

Constantinescu, G., Rieger, J., Winget, M., Paulsen, C., & Seikaly, H. (2017). Patient perception of speech outcomes: The relationship between clinical measures and self-perception of speech function following surgical treatment for oral cancer. *American Journal of Speech-Language Pathology*, *26*. https://doi.org/10.1044/2016_AJSLP-15-0170

Cortes, C., & Vapnik, V. (1995). Support vector network. *Machine Learning*, *20*, 273–297. https://doi.org/10.1007/BF00994018

Dhanuthai, K., Rojanawatsirivej, S., Thosaporn, W., Kintarak, S., Subarnbhesaj, A., Darling, M., Kryshtalskyj, E., Chiang, C., Shin, H., Choi, S.-Y., Lee, S., & Aminishakib, P. (2017). Oral cancer: A multicenter study. *Medicina Oral Patología Oral y Cirugia Bucal*, *23*. https://doi.org/10.4317/medoral.21999

Dibazar, A., Narayanan, S., & Berger, T. (2002). Feature analysis for automatic detection of pathological speech. *Annual International Conference of the IEEE Engineering in Medicine and Biology - Proceedings*, *1*, 182–183 vol.1. https://doi.org/10.1109/IEMBS.2002.1134447

Dibike, Y. B., Velickov, S., Solomatine, D., & Abbott, M. B. (2001). Model induction with support vector machines: Introduction and applications. *Journal of Computing in Civil Engineering*, *15*(3), 208–216. https://doi.org/10.1061/(ASCE)0887-3801(2001)15:3(208)

Dissanayaka, W. L., Pitiyage, G., Kumarasiri, P. V. R., Liyanage, R. L. P. R., Dias, K. D., & Tilakaratne, W. M. (2012). Clinical and histopathologic parameters in survival of oral squamous cell carcinoma. *Oral Surgery, Oral Medicine, Oral Pathology and Oral Radiology*, *113*(4), 518–525. https://doi.org/https://doi.org/10.1016/j.oooo.2011.11.001

Dokuz, Y., & Tufekci, Z. (2021). Mini-batch sample selection strategies for deep learning based speech recognition. *Applied Acoustics*, *171*, 107573. https://doi.org/https://doi.org/10.1016/j.apacoust.2020.107573

Epstein, J., CDA, M., CDA, S., MA, M., & Stevenson-Moore, P. (2001). Quality of life and oral function in patients with radiation therapy for head and neck cancer. *Head & Neck*, *23*, 389–398. https://doi.org/10.1002/hed.1049

Fang, S.-H., Tsao, Y., Hsiao, M.-J., Chen, J.-Y., Lai, Y.-H., Lin, F.-C., & Wang, C.-T. (2018). Detection of pathological voice using cepstrum vectors: A deep learning approach. *Journal of Voice*, *33*. https://doi.org/10.1016/j.jvoice.2018.02.003

Faris, H., Aljarah, I., Habib, M., & Castillo, P. (2020). Hate speech detection using word embedding and deep learning in the arabic language context, 453–460. https://doi.org/10.5220/0008954004530460

Fichaux-Bourin, P., Woisard, V., Grand, S., Puech, M., & Bodin, S. (2009). Validation of a self assessment for speech disorders (phonation handicap index). *Revue de laryngologie - otologie - rhinologie*, *130*, 45–51.

Fraile, R., Saenz-Lechon, N., godino llorente, J., Osma-Ruiz, V., & Fredouille, C. (2009). Automatic detection of laryngeal pathologies in records of sustained vowels by means of mel-frequency cepstral coefficient parameters and differentiation of patients by sex. *Folia phoniatrica et logopaedica : official organ of the International Association of Logopedics and Phoniatrics (IALP)*, *61*, 146–52. https://doi.org/10.1159/000219950

Furia, C. L. B., Kowalski, L. P., do Rosário Dias de Oliveira Latorre, M., Angelis, E. C.-d., Martins, N. M., Barros, A. P., & Ribeiro, K. C. (2001). Speech intelligibility after glossectomy and speech rehabilitation. *Archives of otolaryngology–head & neck surgery*, *127 7*, 877–83.

Godino-Llorente, J., & Gomez-Vilda, P. (2004). Automatic detection of voice impairments by means of short-term cepstral parameters and neural network based detectors. *IEEE Transactions on Biomedical Engineering*, *51*(2), 380–384. https://doi.org/10.1109/TBME.2003.820386

Halpern, B. M., Feng, S., van Son, R., Brekel, M., & Scharenborg, O. (2022a). Low-resource automatic speech recognition and error analyses of oral cancer speech. *Speech Communication*, *141*. https://doi.org/10.1016/j.specom.2022.04.006

Halpern, B. M., Kelly, F., van Son, R., & Alexander, A. (2020b). Residual networks for resisting noise: Analysis of an embeddings-based spoofing countermeasure. https://doi.org/10.21437/Odyssey.2020-46

Halpern, B. M., Rebernik, T., Tienkamp, T., van Son, R., Brekel, M., van den, Wieling, M., Witjes, M., & Scharenborg, O. (2022b). Manipulation of oral cancer speech using neural articulatory synthesis. https://doi.org/10.48550/ARXIV.2203.17072

Halpern, B. M., van Son, R., Brekel, M., & Scharenborg, O. (2020a). Detecting and analysing spontaneous oral cancer speech in the wild, 4826–4830. https://doi.org/10.21437/Interspeech.2020-1598

Hanley, T., Snidecor, J., & Ringel, R. (1966). Some acoustic differences among languages. *Phonetica*, *14*, 97–107. https://doi.org/10.1159/000258520

Hernandez, A., Yeo, E., Kim, S., & Chung, M. (2020). Dysarthria detection and severity assessment using rhythm-based metrics. https://doi.org/10.21437/Interspeech.2020-2354

Hoy, M. B. (2018). Alexa, siri, cortana, and more: An introduction to voice assistants [PMID: 29327988]. *Medical Reference Services Quarterly*, *37*(1), 81–88. https://doi.org/10.1080/02763869.2018.1404391

Huang, D.-Y., Dong, M., & Li, H. (2016). Combining multiple kernel models for automatic intelligibility detection of pathological speech. *2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 6485–6489. https://doi.org/10.1109/ICASSP.2016.7472926

Iannizzotto, G., Lo Bello, L., & Patti, G. (2021). Personal protection equipment detection system for embedded devices based on dnn and fuzzy logic. *Expert Systems with Applications*, *184*, 115447. https://doi.org/https://doi.org/10.1016/j.eswa.2021.115447

Jacobson, B., Johnson, A., Grywalski, C., Silbergleit, A., Jacobson, G., & Benninger, M. (1997). The voice handicap index (vhi): Development and validation. *American Journal of Speech-Language Pathology*, *6*, 66–70.

Joshy, A., & Rajan, R. (2022). Automated dysarthria severity classification: A study on acoustic features and deep learning techniques. *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, *PP*, 1–1. https://doi.org/10.1109/TNSRE.2022.3169814

Kademani, D. (2007). Oral cancer. *Mayo Clinic proceedings*, *82*(7), 878–887. https://doi.org/https://doi.org/10.4065/82.7.878

Kim, H., Jeon, J., Han, Y. J., Joo, Y., Lee, J., Lee, S., & Im, S. (2020). Convolutional neural network classifies pathological voice change in laryngeal cancer with high accuracy. *Journal of Clinical Medicine*, *9*(11). https://doi.org/10.3390/jcm9113415

Kingma, D., & Ba, J. (2014). Adam: A method for stochastic optimization. *International Conference on Learning Representations*.

Kitzing, P. (1986). Ltas criteria pertinent to the measurement of voice quality [Voice Acoustics and Dysphonia Gotland, Sweden, August 1985]. *Journal of Phonetics*, *14*(3), 477–482. https://doi.org/https://doi.org/10.1016/S0095-4470(19)30693-X

Korayem, M., Aljadda, K., & Crandall, D. (2016). Sentiment/subjectivity analysis survey for languages other than english. *Social Network Analysis and Mining , SNAM*, *6*. https://doi.org/10.1007/s13278-016-0381-6

Korpijaakko-Huuhka, A.-M. (1999). Long-lasting speech and oral-motor deficiencies following oral cancer surgery: A retrospective study. *Logopedics Phoniatrics Vocology*, *24*, 97–106. https://doi.org/10.1080/140154399435048

Lai, C.-I., Chen, N., Villalba, J., & Dehak, N. (2019). Assert: Anti-spoofing with squeeze-excitation and residual networks. *arXiv preprint arXiv:1904.01120*.

Lavrentyeva, G., Novoselov, S., Volkova, M., Matveev, Y., & De Marsico, M. (2019). Phonespoof: A new dataset for spoofing attack detection in telephone channel. *ICASSP 2019 - 2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2572–2576. https://doi.org/10.1109/ICASSP.2019.8682942

Lazarus, C., Wall, L., Ward, E., & Yiu, E. (2014). Speech and swallowing following oral, oropharyngeal, and nasopharyngeal cancers.

Li, J., Lavrukhin, V., Ginsburg, B., Leary, R., Kuchaiev, O., Cohen, J. M., Nguyen, H., & Gadde, R. T. (2019). Jasper: An end-to-end convolutional neural acoustic model. https://doi.org/10.48550/ARXIV.1904.03288

Lillicrap, T., Santoro, A., Marris, L., Akerman, C., & Hinton, G. (2020). Backpropagation and the brain. *Nature Reviews Neuroscience*, *21*. https://doi.org/10.1038/s41583-020-0277-3

Mahesh, B. (2018). Machine learning algorithms -a review. *International Journal of Science and Research*, 381–386. https://doi.org/10.21275/ART20203995

Maier, A., Schuster, M., Batliner, A., Noeth, E., & Nkenke, E. (2007). Automatic scoring of the intelligibility in patients with cancer of the oral cavity. *Proceedings of the Annual Conference of the International Speech Communication Association, INTERSPEECH*, *2*, 1206–1209. https://doi.org/10.21437/Interspeech.2007-388

Master, S., De Biase, N., Vieira, V., & Chiari, B. (2006). The long-term average spectrum in research and in the clinical practice of speech therapists. *Pró-fono : revista de atualização científica*, *18*, 111–20. https://doi.org/10.1590/S0104-56872006000100013

Mathog, R. (1991). Rehabilitation of head and neck cancer patients: Consensus on recommendations from the international conference on rehabilitation of the head and neck cancer patient. *Head & Neck*, *13*. https://doi.org/10.1002/hed.2880130102

McFee, B., Raffel, C., Liang, D., Ellis, D., Mcvicar, M., Battenberg, E., & Nieto, O. (2015a). Librosa: Audio and music signal analysis in python, 18–24. https://doi.org/10.25080/Majora-7b98e3ed-003

McFee, B., Raffel, C., Liang, D., Ellis, D., Mcvicar, M., Battenberg, E., & Nieto, O. (2015b). Librosa: Audio and music signal analysis in python, 18–24. https://doi.org/10.25080/Majora-7b98e3ed-003

Minsky, M., & Papert, S. A. (1969). *Perceptrons: An introduction to computational geometry*. The MIT Press.

Mishra, M. (2020). Convolutional neural networks, explained. https://towardsdatascience.com/convolutional-neural-networks-explained-9cc5188c4939

Montavon, G., Binder, A., Lapuschkin, S., Samek, W., & Müller, K.-R. (2019). Layer-wise relevance propagation: An overview. https://doi.org/10.1007/978-3-030-28954-6_10

Mowry, S. E., Ho, A., LoTempio, M. M., Sadeghi, A., Blackwell, K. E., & Wang, M. B. (2006). Quality of life in advanced oropharyngeal carcinoma after chemoradiation versus surgery and radiation. *The Laryngoscope*, *116*(9), 1589–1593. https://doi.org/https://doi.org/10.1097/01.mlg.0000233244.18901.44

Murugan, H. (2020). Speech emotion recognition using cnn. *International Journal of Psychosocial Rehabilitation*, *24*. https://doi.org/10.37200/IJPR/V24I8/PR280260

Nemr, K., Simões-Zenari, M., Cordeiro, G. F., Tsuji, D., Ogawa, A. I., Ubrig, M. T., & Menezes, M. H. M. (2012). Grbas and cape-v scales: High reliability and consensus when applied at different times. *Journal of Voice*, *26*(6), 812.e17–812.e22. https://doi.org/https://doi.org/10.1016/j.jvoice.2012.03.005

Nicoletti, G., Soutar, D., Jackson, M., Wrench, A., Robertson, G., & Robertson, C. (2004). Objective assessment of speech after surgical treatment for oral cancer: Experience from 196 selected cases. *Plastic and Reconstructive Surgery*, *113*(1), 114–125. https://doi.org/10.1097/01.PRS.0000095937.45812.84

Orozco, J. R., Hoenig, F., Arias-Londoño, J. D., Vargas-Bonilla, J., Daqrouq, K., Skodda, S., Rusz, J., & Noeth, E. (2016). Automatic detection of parkinson's disease in running speech spoken in three different languages. *The Journal of the Acoustical Society of America*, *139*, 481–500. https://doi.org/10.1121/1.4939739

Pace-balzan, A., Shaw, R., & Butterworth, C. (2011). Oral rehabilitation following treatment for oral cancer. *Periodontology 2000*, *57*, 102–17. https://doi.org/10.1111/j.1600-0757.2011.00384.x

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., & Duchesnay, E. (2011). Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, *12*, 2825–2830.

Phillips, J. (2021). Gradient descent. https://doi.org/10.1007/978-3-030-62341-8_6

Plaza-del-Arco, F. M., Molina-González, M. D., Ureña-López, L. A., & Martín-Valdivia, M. T. (2021). Comparing pre-trained language models for spanish hate speech detection. *Expert Systems with Applications*, *166*, 114120. https://doi.org/https://doi.org/10.1016/j.eswa.2020.114120

Povey, D., Ghoshal, A., Boulianne, G., Burget, L., Glembek, O., Goel, N., Hannemann, M., Motlíček, P., Qian, Y., Schwarz, P., Silovský, J., Stemmer, G., & Vesel, K. (2011). The kaldi speech recognition toolkit. *IEEE 2011 Workshop on Automatic Speech Recognition and Understanding*.

Prajapati, G. L., & Patle, A. (2010). On performing classification using svm with radial basis and polynomial kernel functions. *2010 3rd International Conference on Emerging Trends in Engineering and Technology*, 512–515. https://doi.org/10.1109/ICETET.2010.134

Qian, Y., Bi, M., Tan, T., & Yu, K. (2016). Very deep convolutional neural networks for noise robust speech recognition. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, *24*(12), 2263–2276. https://doi.org/10.1109/TASLP.2016.2602884

Ramoo, D. (2021). *Psychology of language (open education resource)*. https://opentextbc.ca/psyclanguage/

Ravanelli, M., & Bengio, Y. (2018). Speaker recognition from raw waveform with sincnet. https://doi.org/10.48550/ARXIV.1808.00158

Riedhammer, K., Bocklet, T., Ghoshal, A., & Povey, D. (2012). Revisiting semi-continuous hidden markov models. *2012 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 4721–4724. https://doi.org/10.1109/ICASSP.2012.6288973

Rieger, J., Happonen, R.-P., Harris, J., & Seikaly, H. (2010). Speech after radial forearm free flap reconstruction of the tongue: A longitudinal acoustic study of vowel and diphthong sounds. *Clinical linguistics & phonetics*, *24*, 41–54. https://doi.org/10.3109/02699200903340758

Rinkel, R., Leeuw, I., Reij, E., Aaronson, N., & Leemans, C. (2008). Speech handicap index in patients with oral and pharyngeal cancer: Better understanding of patients' complaints. *Head & Neck*, *30*, 868–874. https://doi.org/10.1002/hed.20795

Rivera, C. (2015). Essentials of oral cancer. *International journal of clinical and experimental pathology*, *8*, 11884–11894. https://doi.org/10.5281/zenodo.192487

Rosenblatt, F. (1960). Perceptron simulation experiments. *Proceedings of the IRE*, *48*(3), 301–309. https://doi.org/10.1109/JRPROC.1960.287598

Saravanan, G., Ranganathan, V., Gandhi, A., & Jaya, V. (2016). Speech outcome in oral cancer patients – pre- and post-operative evaluation: A cross-sectional study. *Indian Journal of Palliative Care*, *22*, 499–503. https://doi.org/10.4103/0973-1075.191858

Sharma, G., Umapathy, K., & Krishnan, S. (2020). Trends in audio signal feature extraction methods. *Applied Acoustics*, *158*, 107020. https://doi.org/10.1016/j.apacoust.2019.107020

Shellikeri, S., Green, J., Kulkarni, M., Rong, P., Martino, R., Zinman, L., & Yunusova, Y. (2016). Speech movement measures as markers of bulbar disease in amyotrophic lateral sclerosis. *Journal of Speech Language and Hearing Research*, *59*, 1. https://doi.org/10.1044/2016_JSLHR-S-15-0238

Shelton, I., & Garves, M. (1985). Use of visual techniques in therapy for developmental apraxia of speech. *Language Speech and Hearing Services in Schools*, *16*, 129. https://doi.org/10.1044/0161-1461.1602.129

Shield, K. D., Ferlay, J., Jemal, A., Sankaranarayanan, R., Chaturvedi, A. K., Bray, F., & Soerjomataram, I. (2017). The global incidence of lip, oral cavity, and pharyngeal cancers by subsite in 2012. *CA: A Cancer Journal for Clinicians*, *67*(1), 51–64. https://doi.org/https://doi.org/10.3322/caac.21384

Smith, L., & Goberman, A. (2014). Long-time average spectrum in individuals with parkinson disease. *NeuroRehabilitation*, *35*. https://doi.org/10.3233/NRE-141102

Stemmer, G. (2005). Modelling variability in speech recognition.

Stipancic, K., Palmer, K., Rowe, H., Yunusova, Y., Berry, J., & Green, J. (2021). "you say severe, i say mild": Toward an empirical classification of dysarthria severity. *Journal of Speech, Language, and Hearing Research*, *64*, 1–18. https://doi.org/10.1044/2021_JSLHR-21-00197

Suárez-Cunqueiro, M. M., Schramm, A., Schoen, R., Seoane-Lestón, J., Otero-Cepeda, X. L., Bormann, K.-H., Kokemueller, H., Metzger, M. C., Diz-Dios, P., & Gellrich, N.-C. (2008). Speech and swallowing impairment after treatment for oral and oropharyngeal cancer. *Archives of otolaryngology–head & neck surgery*, *134 12*, 1299–304.

Sussman, J., & Tjaden, K. (2012). Perceptual measures of speech from individuals with parkinson's disease and multiple sclerosis: Intelligibility and beyond. *Journal of speech, language, and hearing research : JSLHR*, *55*, 1208–19. https://doi.org/10.1044/1092-4388(2011/11-0048)

Tanner, K., Roy, N., Ash, A., & Buder, E. (2005). Spectral moments of the long-term average spectrum: Sensitive indices of voice change after therapy? *Journal of voice : official journal of the Voice Foundation*, *19*, 211–222. https://doi.org/10.1016/j.jvoice.2004.02.005

Tjaden, K., Sussman, J., & Wilding, G. (2014). Impact of clear, loud, and slow speech on scaled intelligibility and speech severity in parkinson's disease and multiple sclerosis. *Journal of speech, language, and hearing research : JSLHR*, *57*. https://doi.org/10.1044/2014_JSLHR-S-12-0372

Tripathi, A., Bhosale, S., & Kopparapu, S. K. (2020). Improved speaker independent dysarthria intelligibility classification using deepspeech posteriors, 6114–6118. https://doi.org/10.1109/ICASSP40776.2020.9054492

Van den Steen, L., Nuffelen, G., Guns, C., Groote, M., Pinson, L., & Bodt, M. (2011). De spraak handicap index: Een instrument voor zelfevaluatie by dysartriepatienten. *Logopedie*, *24*, 26–30.

Vandeghinste, V., Bulté, B., & Augustinus, L. (2019). Wablieft: An easy-to-read newspaper corpus for dutch.

van Son, R., Middag, C., & Demuynck, K. (2018). Vowel space as a tool to evaluate articulation problems, 357–361. https://doi.org/10.21437/Interspeech.2018-68

Windrich, M., Maier, A., Kohler, R., Nöth, E., Nkenke, E., Eysholdt, U., & Schuster, M. (2008). Automatic quantification of speech intelligibility of adults with oral squamous cell carcinoma. *Folia Phoniatrica et Logopaedica*, *60*(3), 151–156. https://doi.org/10.1159/000121004

Woisard, V., Balaguer, M., Fredouille, C., Farinas, J., Ghio, A., Lalain, M., Puech, M., Astésano, C., Pinquier, J., & Lepage, B. (2021). Construction of an automatic score for the evaluation of speech disorders among patients treated for a cancer of the oral cavity or the oropharynx: The carcinologic speech severity index. *Head & Neck*, *44*, 71–88.

Xue, W., Hout, R., Boogmans, F., Ganzeboom, M., Cucchiarini, C., & Strik, H. (2021). Speech intelligibility of dysarthric speech: Human scores and acoustic-phonetic features. https://doi.org/10.21437/Interspeech.2021-1189

Yunusova, Y., Graham, N., Shellikeri, S., Phuong, K., Kulkarni, M., Rochon, E., Tang-Wai, D., Chow, T., Black, S., Zinman, L., & Green, J. (2016). Profiling speech and pausing in amyotrophic lateral sclerosis (als) and frontotemporal dementia (ftd). *PloS one*, *11*, e0147573. https://doi.org/10.1371/journal.pone.0147573

Zai. (2021). Logistic regression explained. https://towardsdatascience.com/logistic-regression-explained-9ee73cede081

Zhang, Z. (2019). Support vector machine explained. https://towardsdatascience.com/support-vector-machine-explained-8bfef2f17e71

Zimmermann, A., Sader, R., Hoole, P., Bressmann, T., Mady, K., & Horch, H.-H. (2003). The influence of oral cavity tumour treatment on the voice quality and on fundamental frequency. *Clinical linguistics & phonetics*, *17*, 273–81. https://doi.org/10.1080/0269920031000080073

Appendices

A Overview of OC TNM Staging

Table 13: *TNM staging for OC as presented by Cancernet (2021) and Kademani (2007). PT refers to the primary tumor, T refers to tumor, NS refers to the nodal status, N refers to node, LN refers to the lymph nodes, M refers to metastasis and DM refers to distant metastasis. Please be aware that there are different types of TNM staging depending on the type of cancer. In this case, we only presented the staging for OC.*

PT	***Explanation***	***NS***	***Explanation***
TX	PT cannot be evaluated	*NX*	Regional LN cannot be evaluated
T1	PT <2 cm	*N1*	M to single ipsilateral LN; <3cm
T2	PT 2-4cm	*N2a*	M to single ipsilateral LN; 3-6cm
T3	PT 4-10cm	*N2b*	M to various ipsilateral LN; <6cm
T4	PT >10cm	*N2c*	M to bilateral/contralateral LN; <6cm
Staging	***Explanation***	*N3*	M to any LN; >6cm
Stage 1	T1, N0, M0	***DM***	***Explanation***
Stage 2	T2, N0, M0	*Mx*	M cannot be evaluated
Stage 3	T3N0M0; T1N1M0; T2N1M0; T3N1M0	*M0*	No spreading of cancer to other body parts
Stage 4	Any T4, N2, N3 or M1 lesion	*M1*	Spreading of cancer to other body parts

B SHI: Original Version by Rinkel et al. (2008)

Table 1. Speech Handicap Index form.

Item
1. My speech makes it difficult for people to understand me
2. I run out of air when I speak
3. The intelligibility of my speech varies throughout the day
4. My speech makes me feel incompetent
5. People ask me why I'm hard to understand
6. I feel annoyed when people ask me to repeat
7. I avoid using the phone
8. I'm tense when talking to others because of my speech
9. My articulation is unclear
10. People have difficulty understanding me in a noisy room
11. I tend to avoid groups of people because of my speech
12. People seem irritated with my speech
13. People ask me to repeat myself when speaking face-to-face.
14. I speak with friends and neighbors or relatives less often because of my speech
15. I feel as though I have to strain to speak
16. I find other people don't understand my speaking problem
17. My speaking difficulties restrict my personal and social life
18. The intelligibility is unpredictable
19. I feel left out of conversations because of my speech
20. I use a great deal of effort to speak
21. My speech is worse in the evening
22. My speech problem causes me to lose income
23. I try to change my speech to sound different
24. My speech problem upsets me
25. I am less outgoing because of my speech problem
26. My family has difficulty understanding me when I call them throughout the house
27. My speech makes me feel handicapped
28. I have difficulties to continue a conversation because of my speech
29. I feel embarrassed when people ask me to repeat
30. I'm ashamed of my speech problem

How do you rate your own speech at this moment (please circle the right answer)? Excellent Good Average Bad

C SHI: Dutch Adaptation by Van den Steen et al. (2011)

SPRAAK HANDICAP INDEX

Naam patiënt: ..... Datum: .....

Geboortedatum: .....

Type dysartrie: ..... Score: P ..../20

Etiologie: ..... F ..../20

Datum ontstaan: ..... E ..../20

Ernstgraad dysartrie: *licht – matig – ernstig* Totaal: ...../60

Dit zijn beweringen die veel mensen gebruikt hebben om hun spraak en de gevolgen van hun spraak op hun leven te beschrijven. Zet een kruisje bij dat antwoord dat aangeeft hoe dikwijls u dezelfde ervaring heeft.

		NOOIT	BIJNA NOOIT	SOMS	BIJNA ALTIJD	ALTIJD
P1	De snelheid waarmee ik praat is veranderd.					
P2	Ik heb het moeilijk om met mijn stem emoties uit te drukken.					
P3	Ik heb moeilijkheden om goed te articuleren wanneer ik praat.					
P4	Ik moet een inspanning doen om te praten.					
P5	Ik ben buiten adem als ik praat.					
F1	Ik heb het moeilijk om mondeling uit te drukken wat ik nodig heb (eten, drinken, WC,...).					
F2	Ik schaam me om mijn gedachten en ideeën uit te drukken.					
F3	Ik heb het moeilijk om te communiceren met mensen die ik niet goed ken.					
F4	Door mijn spraakprobleem vraagt men mij vaak om iets te herhalen.					
F5	Ik vermijd gesprekken met mijn familie, vrienden, buren.					
E1	Ik lijd onder mijn manier van praten.					
E2	Mijn spraakmoeilijkheden beperken mijn persoonlijk en sociaal leven.					
E3	Ik vind dat anderen mijn spraakproblemen niet begrijpen.					
E4	Mensen lijken geïrriteerd door mijn spraakproblemen.					
E5	Ik voel me gehandicapt omwille van mijn spraakmoeilijkheden.					

Abbreviations: Questions with indicator *P* refer to the physical impact, those with indicator *F* refer to the functional impact and those with indicator *E* refer to the emotional impact of a patient's speech on their quality of life.

D Translated SHI Questions from Van den Steen et al. (2011)

1. P1: The speed with which I speak has changed.
2. P2: I find it difficult to express my emotions through my voice.
3. P3: I have difficulty articulating well when I speak.
4. P4: I have to put in a lot of effort when I speak.
5. P5: I am out of breath when I speak.
6. F1: I find it hard to express orally what I need [food, drinks, bathroom, ...].
7. F2: The thought of expressing my thoughts and ideas embarrasses me.
8. F3: I find it difficult to communicate with people whom I don't know well.
9. F4: Due to my speech impairment people often ask me to repeat something.
10. F5: I avoid conversations with my family, friends, neighbors.
11. E1: I suffer because of the manner in which I speak.
12. E2: My speech impairments restrict me in my personal and social life.
13. E3: I feel like others don't understand my speech impairments.
14. E4: People seem irritated by my speech impairments.
15. E5: I feel disabled as a result of my speech difficulties.

E Overview of All Participants with Their Corresponding Speaker ID, Group and Sex

Table 14: *Overview of the participants included in the speech dataset. Sex of speaker is indicated either male (M) or female (F). Healthy controls (HC) and OC patients (PT) all received a speaker ID based on their group (CON for control speaker and OC for oral cancer speaker).*

ID	***Group***	***Sex***
HC1	CON	M
HC2	CON	M
HC3	CON	F
HC4	CON	F
HC5	CON	F
PT1	OC	M
PT2	OC	M
PT3	OC	M
PT4	OC	F
PT5	OC	F
PT6	OC	F

F Complete Stimuli Set without Repetitions

Custom sentences (excluded)

1. Hij heeft tamme shock gezegd.
2. Hij heeft tamme sock gezegd.
3. Hij heeft tamme biet gezegd.
4. Hij heeft tamme boet gezegd.
5. Hij heeft tamme baat gezegd.

Papa en Marloes

1. Papa en Marloes staan op het station.
2. Ze wachten op de trein.
3. Eerst hebben ze een kaartje gekocht.
4. Er stond een hele lange rij, dus dat duurde wel even.
5. Nu wachten ze tot de trein eraan komt.
6. Het is al vijf over drie, dus het duurt nog vier minuten.
7. Er staan nog veel meer mensen te wachten.
8. Marloes kijkt naar links, in de verte ziet ze de trein al aankomen.

Man uit Finland

1. Er was eens een man uit Finland.
2. Hij had veel geld gespaard.
3. Dat was voor de auto van zijn dromen.
4. Hij nam de trein om de auto te gaan kopen.
5. Maar de man was bang voor dieven.
6. Hij bewaarde het geld in zijn onderbroek.
7. Hij droomde al van de eerste rit in de nieuwe wagen.
8. Plots moest hij naar het toilet.
9. De man dacht niet meer aan het geld.
10. Het zakje met geld viel recht in de pot.
11. En de man spoelde door.
12. Daar ging zijn fraaie plan!
13. Gelukkig was de politie in de buurt.
14. Die vond het zakje terug op het spoor.

Noordenwind en de zon

1. De noordenwind en de zon waren erover aan het redetwisten wie de sterkste was van hen beiden.
2. Juist op dat moment kwam er een reiziger aan, die gehuld was in een warme mantel.
3. Ze waren het erover eens dat degene die er als eerste in slaagde de reiziger zijn mantel uit te doen, als sterker moest worden beschouwd dan de ander.
4. De noordenwind begon toen uit alle macht te blazen.

5. Maar hoe harder hij blies, des te dichter trok de reiziger zijn mantel om zich heen.
6. Ten lange leste gaf de noordenwind het op.
7. Daarna begon de zon krachtig te stralen, en hierop trok de reiziger onmiddellijk zijn mantel uit.
8. De noordenwind moest dus wel bekennen dat de zon van hen beiden de sterkste was.

Els gaat naar markt

1. Het is zaterdag.
2. Els heeft vrij.
3. Ze loopt door de stad.
4. Het is prachtig weer, de lucht is blauw.
5. Op straat ziet ze Bart op de fiets.
6. Hij wacht voor het rode licht.
7. Als Bart haar ziet, zwaait hij.
8. Els loopt weer verder.
9. Bij de bakker koopt ze brood, bij de slager koopt ze vlees.
10. Als het vijf uur is gaat ze terug, zodat ze op tijd weer thuis is.

Meneer van Dam

1. Vanmorgen ging meneer van Dam naar de groenteman.
2. Namelijk om een mand mandarijnen te kopen.
3. Aan zijn arm nam hij een mand mee om de mandarijnen in te doen.
4. Na een minuut of tien stond meneer van Dam in de winkel.
5. En hij nam een mand mandarijnen mee en ook maar meteen negen bananen en een mooie ananas.
6. Met zijn mand aan zijn arm ging hij toen snel naar huis.

Jorinde en Joringel

1. Er was eens een oud kasteel midden in een diep en donker bos.
2. Daarin woonde een oude heks helemaal alleen.
3. Overdag veranderde ze zich in een kat of een uil, maar ‘s avonds werd ze weer een mens.
4. Ze kon dieren en vogels naar zich toe lokken.
5. Die dieren slachtte, kookte en braadde ze dan.
6. Wanneer iemand binnen honderd meter van het kasteel kwam, moest hij stilstaan en kon zich niet meer verroeren.
7. Dit duurde totdat de heks hem met een spreuk verloste.
8. Wanneer er echter een onschuldig meisje te dicht bij haar kasteel kwam, veranderde de heks haar in een vogel en sloot haar op in een kooitje.
9. Dat kooitje bracht ze dan naar een zaal van haar kasteel.
10. Ze had wel zevenduizend kooien met zulke bijzondere vogels in haar kasteel.
11. Nu was er eens een meisje dat Jorinde heette.
12. Ze was mooier dan alle andere meisjes en was verloofd met de knappe Joringel.

13. Ze zouden over een paar dagen gaan trouwen en ze hadden veel plezier met elkaar.
14. Om eens rustig samen te kunnen praten, gingen ze in het bos wandelen.
15. ‘Pas op’, zei Joringel, ‘dat je niet te dicht bij het kasteel komt’.
16. Het was een mooie avond.
17. Het heldere zonlicht scheen tussen de boomstammen door in het donkere groen van het bos.
18. De tortelduif zong klagelijk in de oude beuk.
19. Jorinde huilde een beetje.
20. Ze ging in de zon zitten en klaagde.
21. Joringel klaagde ook.
22. Ze waren verdrietig, alsof ze moesten sterven.
23. Ze keken om zich heen en waren verdwaald.
24. Ze wisten niet meer hoe ze thuis moesten komen.
25. De zon stond nog maar half boven de berg en voor de helft was ze al onder.
26. Joringel keek door de struiken en zag vlakbij de oude muur van het kasteel.
27. Hij schrok en werd doodsbang.
28. Jorinde zong:
29. Mijn vogeltje met het rode ringetje
30. Zingt lijden, lijden, lijden:
31. Het zingt voor het duifje, zingt voor zijn dood,
32. Zingt lijden, lij, twiet, twiet, twiet.
33. Joringel keek naar Jorinde.
34. Jorinde was in een nachtegaal veranderd die twiet, twiet zong.
35. Een uil met gloeiende ogen vloog drie keer om hen heen en schreeuwde drie keer oehoe, oehoe, oehoe.
36. Joringel kon zich niet meer bewegen.
37. Hij stond erbij als van steen, kon niet huilen, niet praten, geen hand of voet bewegen.
38. Nu was de zon ondergegaan.
39. De uil vloog in een struik en direct kwam er een kromme, oude vrouw tevoorschijn.
40. Ze was geel en mager.
41. Ze had grote rode ogen en een kromme neus die met de punt tot aan haar kin kwam.
42. Ze mompelde wat, ving de nachtegaal en droeg die in haar hand weg.
43. Joringel kon niets zeggen, niet van z’n plaats komen.
44. De nachtegaal was weg.
45. Eindelijk kwam de oude vrouw terug en zei met een doffe stem:
46. ‘Gegroet Zachiël’
47. Maak los, op het juiste moment, wanneer het maantje in het kooitje schijnt.
48. Toen was Joringel verlost.
49. Hij viel voor de oude vrouw op de knieën en smeekte haar om hem Jorinde terug te geven.
50. Maar ze zei dat hij Jorinde nooit meer terug zou krijgen en ging weg.

51. Hij riep, hij huilde, hij jammerde, maar het was allemaal voor niets.
52. ‘Oh, wat moet er van mij worden?’ Joringel ging weg en kwam uiteindelijk in een vreemd dorp.
53. Daar hoedde hij lange tijd de schapen.
54. Vaak liep hij rond het kasteel, maar hij kwam nooit te dichtbij.
55. Een keer droomde hij ‘s nachts dat hij een bloedrode bloem vond met in het midden een prachtige grote parel.
56. Hij plukte de bloem en ging ermee naar het kasteel.
57. Alles wat hij met de bloem aanraakte werd van de betovering bevrijd.
58. Ook droomde hij dat hij daardoor zijn Jorinde teruggekregen had.
59. ‘s Morgens, nadat hij wakker werd, begon hij door berg en dal naar zo’n bloem te zoeken.
60. Hij zocht tot aan de negende dag.
61. Toen vond hij de bloem in de vroege ochtend.
62. In het midden lag een grote dauwdruppel, zo groot als de mooiste parel.
63. Joringel liep dag en nacht en droeg de bloem naar het kasteel.
64. Toen hij dichtbij het kasteel gekomen was, verstijfde hij niet, maar hij liep door tot aan de deur.
65. Joringel werd heel blij, raakte de deur aan met de bloem en de deur sprong open.
66. Joringel ging naar binnen, liep over de binnenplaats en luisterde goed of hij de vele vogels kon horen.
67. Toen hoorde hij ze fluiten.
68. Hij liep in de richting van het gefluit en vond de zaal.
69. Daar was de heks bezig de vogels in hun zevenduizend kooien te voeren.
70. Toen ze Joringel zag werd ze kwaad, heel erg kwaad.
71. Ze schold, tierde en spuwde gif en gal naar hem.
72. Maar ze kon niet bij hem in de buurt komen.
73. Joringel lette niet op haar en bekeek de kooien met de vogels.
74. Er waren vele honderden nachtegalen, hoe moest hij nou Jorinde terugvinden?
75. Toen hij zo rondkeek, merkte hij, dat de oude vrouw stiekem een vogelkooitje wegpakte en daarmee naar de deur liep.
76. Snel sprong hij erheen en raakte het kooitje en de oude vrouw aan met de bloem.
77. Nu kon de heks niet meer toveren, en Jorinde stond weer voor hem.
78. Ze vloog hem om de hals en was zo mooi als vroeger.
79. Daarna veranderde hij ook alle andere vogels weer in meisjes en ging met zijn Jorinde naar huis.
80. En ze leefden nog lang en gelukkig met elkaar.

Wablieft (news source)

1. Het concert mocht doorgaan, maar zonder licht of decor!
2. Lachgas is gevaarlijk.
3. Voor beide surfers stuurden de hulpdiensten ziekenwagens en reddingsboten uit.
4. Een e-boek is altijd goedkoper dan hetzelfde boek op papier.

5. Op dinsdag 10 oktober voetbalt België in Brussel tegen Cyprus.
6. Sindsdien vond de tocht al 15 keer plaats.
7. Op zondag 8 september is het feest.
8. Kenners noemen Messi de beste voetballer ter wereld.
9. Samba is de meest bekende muzieksoort uit Brazilië.
10. Die vond plaats op woensdag 30 oktober.
11. Ik ben Hank, steward voor de passagiers in de tweede klasse.
12. De pikante hamburger uit Bristol kost 30 euro.
13. PepsiCo is het bedrijf achter frisdrank Pepsi.
14. Bangkok is de hoofdstad van Thailand in Azië.
15. De Nederlandse burgers kiezen op 12 september een nieuwe regering.
16. Door haar bekendheid kreeg Moeder Teresa miljoenen euro's van schenkers.
17. Alle Cyprioten zouden een hoge taks betalen op hun spaargeld.
18. Facebook onthoudt welke websites de gebruikers nog bezoeken.
19. De cursisten spraken op vrijdag 7 september met de politici.
20. Hij bezat de Europese titel sinds de zomer van 2014.
21. Dat is de belangrijkste rechtbank van het land.
22. Sterke lopers onder de veldrijders zagen hun kans.
23. De officiële resultaten zijn waarschijnlijk morgen, donderdag, bekend.
24. Arbeiders sloopten stukken van de tempel met bulldozers.
25. De Warmathon hoopt duizenden mensen op straat te krijgen.
26. De ziekenfondsen betalen sinds 2016 het remgeld terug voor kinderen.
27. De meeste pastoors zijn niet tevreden over aartsbisschop Léonard.
28. Vele tienduizenden mensen bekijken hun filmpjes op de website YouTube.
29. Met Pasen was minder dan één op vijf hotelkamers bezet.
30. De chefs bij Noma koken met producten uit de streek.
31. Dat vindt plaats op 25 september in Kopenhagen in Denemarken.
32. De pakjes brengt hij pas op 6 december.
33. Er zijn wedstrijden voor de best verklede bezoekers.
34. Op 6 december komt Sinterklaas langs.
35. Behalve in Brazilië, daar spreken mensen Portugees.
36. Voor de quizploeg probeert hij alles te onthouden.
37. De capsule in Boston zat er sinds 1914.
38. Eén straat heeft bijzondere parkeermeters.
39. Je hebt ook de sociale netwerken op internet, zoals Facebook.
40. Uiteindelijk bleken de toeschouwers toch in 'veilige' zones te staan.
41. Twee bedrijven uit Italië maken samen pasta.
42. Enkel president Obama kan de pijplijn nog tegenhouden.
43. In Groot-Brittannië vond het wereldkampioenschap darts plaats.

44. Na het wereldkampioenschap in Brazilië wilden ze snel naar huis.
45. Dit betekent net hetzelfde als keuze 2.
46. Die bleek 18 keer sterker dan eerst gedacht.
47. Haar tegenstanders blijven steken op 24 zetels.
48. Toen vond in Brazilië het wereldkampioenschap voetbal plaats.
49. De bibliotheek heeft 20 jaar lang cd's gekocht.
50. Moeder Teresa wordt op 4 september heilig verklaard.
51. Australië lijdt onder de zwaarste bosbranden sinds jaren.
52. Hij was 35 jaar sportjournalist op de radio.
53. In België is dat verboden op tijdelijke plaatsen.
54. Je kan dat tijdelijk gratis beluisteren op iTunes.
55. Dat zei de Amerikaanse president Obama op tv.
56. Tijdens het bezoek was er protest tegen Obama.
57. Dat jaar kwamen de eerste 600 bezoekers naar het park.
58. Elektronische maaltijdcheques kosten veel minder dan papieren cheques.
59. In 2013 stopt hij ook als president van China.
60. Dat was de zesde rally voor het Belgisch kampioenschap.
61. Op dit moment zijn er 800 strips beschikbaar.
62. Zondag kwamen de Europese ministers van Financiën samen in Luxemburg.
63. Belgische organisaties gebruikten 6 miljoen voor noodhulp.
64. Tanken langs de snelweg blijft heel duur.
65. Het decor drijft op het Bodenmeer.
66. Dat is dé auto in Oost-Duitsland.
67. De Turkse president Erdogan sprak het land toe.
68. De onderzoekers zetten nu aardbeiplantjes op duizend vensterbanken.
69. De spelers hadden achteraf kritiek op trainer Weiler.
70. Ook België heeft redders en dokters ter plaatse.
71. De eerste voorstelling vindt plaats op zondag 20 september.
72. De Britse zangeres Adele is met succes geopereerd.
73. De rechtbank bestaat sinds 2002.
74. Zodra de index 2 procent stijgt, helpt de overheid.
75. Bijvoorbeeld de presidenten van Rusland, China en Syrië.
76. Het gaat bijvoorbeeld om kwetsende opmerkingen op Facebook.