university of groningen / campus fryslân

Automatic speech recognition and error analyses of Dutch oral cancer speech

Kirsten Wildenburg

university of groningen / campus fryslân

University of Groningen

Automatic speech recognition and error analyses of Dutch oral cancer speech

Master's Thesis

To fulfill the requirements for the degree of
Master of Science
at University of Groningen under the supervision of
Dr. V. Verkhodanova (University of Groningen)
and
Bence M. Halpern (Netherlands Cancer Institute, University of Amsterdam)

Kirsten Wildenburg (s3468968)

July 10, 2022

Acknowledgments

First, I would like to express my profound gratitude and appreciation to my external supervisor Bence Halpern, without whom I would not have been able to conduct this research. He patiently answered all of my questions and provided me with invaluable help and feedback every step of the way.

Secondly, I would like to sincerely thank my supervisor, Dr. Vass Verkhodanova. She gave me insightful feedback, ideas, and most of all emotional support. Her enthusiasm was contagious throughout this project.

Thirdly, I would like to gratefully acknowledge the opportunity given to me by the creators of the oral cancer speech dataset to work with this corpus and use it to perform experiments.

Lastly, I would like to wholeheartedly thank my friend and fellow master student Janay Monen. We had many helpful and interesting discussions on our theses, but I mostly want to thank her for the support throughout this roller coaster of a project.

Abstract

Approximately 500.000 people are diagnosed with oral cancer yearly (Shield et al., 2016), and the treatment of oral cancer often leads to impaired speech intelligibility (Lazarus et al., 2014). Automatic speech recognition (ASR) systems could ameliorate oral cancer survivors' quality of life since it could ease their communication and it could also be applied clinically (Windrich et al., 2008). Therefore, this study aims to investigate what phonemes cause higher recognition error rates in standard end-to-end (E2E) ASR systems for oral cancer speech compared to healthy speech in Dutch, as well as the influence of the surgical treatment on the ASR performance. We use the ESPnet E2E ASR system that adopts a hybrid CTC-attention architecture in combination with a Conformer model that was pre-trained on the CGN corpus containing healthy Dutch speech. After running our Dutch oral cancer speech dataset through the ASR system, we perform an extensive error analysis on both the phoneme and articulatory feature level. In agreement with the literature (e.g. Halpern et al., 2022), our results reveal that the E2E ASR system performs significantly poorer for oral cancer speech than for healthy speech. Especially the production of /k/ elicits higher recognition error rates in oral cancer speech, which is in line with previous research (e.g. Borggreven et al., 2005; de Bruijn et al., 2009). Our articulatory feature analysis supports these findings as it shows that velar consonants are the second most challenging articulatory feature class to be recognized in oral cancer speech, and that plosives are misrecognized most frequently by the ASR system in terms of manner of articulation. Although previous studies report on sibilants being misrecognized in oral cancer speech (e.g. Laaksonen et al., 2011), our results do not show sibilants to be more challenging for the ASR system to capture in oral cancer speech, which is in accordance with the findings of Halpern et al. (2022). In addition, the speech of oral cancer patients who underwent a mandibulectomy seems to obtain higher recognition error rates than the speech of patients who underwent a (partial) glossectomy, although the difference between WERs fails to reach significance. The findings of this study contribute to the development of Dutch ASR systems for oral cancer speech.

Contents

Page

List of Figures **7**

List of Tables **8**

1 Introduction **9**
1.1 Research questions and hypotheses . . . 10
1.2 Thesis Outline . . . 10

2 Background Literature **11**
2.1 Failure of speech production . . . 11
2.2 Oral cancer speech . . . 12
2.2.1 Plosives . . . 12
2.2.2 (Alveolar) sibilants . . . 13
2.2.3 Vowels . . . 13
2.3 Deep Learning (DL) . . . 14
2.4 Automatic Speech Recognition (ASR) . . . 15
2.4.1 End-to-end ASR . . . 16
2.5 ASR for pathological speech . . . 20
2.5.1 ASR for oral cancer speech . . . 20

3 Methodology **22**
3.1 Dataset . . . 22
3.1.1 Oral cancer speech dataset . . . 22
3.1.2 CGN corpus . . . 23
3.2 Model . . . 23
3.2.1 End-to-end ASR: ESPnet . . . 23
3.3 ASR evaluation: error analyses . . . 24
3.3.1 Word error rate (WER) . . . 24
3.3.2 Phoneme error analysis and articulatory feature error analysis . . . 25

4 Results **26**
4.1 Word Error Rate . . . 26
4.1.1 Effect of surgical treatment on the WER . . . 26
4.2 Phoneme error analysis . . . 27
4.2.1 Effect of surgical treatment on the PER . . . 28
4.3 Articulatory feature error analysis . . . 29
4.3.1 Place of Articulation (PoA) . . . 29
4.3.2 Manner of Articulation (MoA) . . . 29
4.3.3 Effect of surgical treatment on the AFER . . . 30

5 Discussion **32**
5.1 Recognition errors in healthy and oral cancer speech . . . 32
5.2 Recognition errors specific to oral cancer speech . . . 32
5.3 Influence of the type of surgical treatment . . . 34

5.4 Limitations and future recommendations 36

6 Conclusion **37**

Bibliography **39**

Appendices **45**

A Data agreement **45**

B Sentences read by participants **46**

C PER results per participant **54**

List of Figures

1 Source-filter theory of speech production. 11
2 Structure of a neural network 15
3 Traditional ASR pipeline 16
4 Structure of a RNN 18
5 Attention-based alignments. 18
6 Standard ESPnet recipe. 19
7 WER (%) results per surgical treatment 26
8 PER (%) results 28
9 PER (%) results per surgical treatment 28
10 AFER (%) results for PoA per surgical treatment 30
11 AFER (%) results for MoA per surgical treatment 31

List of Tables

1 Participant characteristics. 22
2 Phonemes in Dutch. 25
3 Word recognition errors. 27
4 Recognition errors on the articulatory feature level. 29
5 Recognition errors on the phoneme level for consonants. 54
6 Recognition errors on the phoneme level for vowels. 55

1 Introduction

Over the last couple of years automatic speech recognition (ASR) has significantly improved, and its numerous applications to make people's lives more convenient range from voice assistants (such as Siri and Alexa) to voice interaction systems in banks and hospitals. The improved performance of ASR systems is the result of the introduction of deep learning (Graves and Jaitly, 2014). Consequently, current ASR systems need large amounts of input data to be trained on, as the systems learn how to recognize speech from this data (Alzubaidi et al., 2021). State-of-the-art ASR systems work very well for speech that is similar to the training data, which usually comprises native speech of adult speakers who have a standardized dialect, without a speech impairment. However, ASR systems do not work well for people whose speech diverges from standard speech (e.g. Muhammad et al., 2011; Tatman and Kasten, 2017; Koenecke et al., 2020), even though these people could arguably benefit the most from ASR systems (e.g. Windrich et al., 2008). The speech of people who have been diagnosed with and treated for oral cancer falls within this group.

Globally, approximately 529.500 people have to battle oral cancer every year (Shield et al., 2016). Survivors of this disease can have problems affecting several basic functions such as swallowing (Lazarus et al., 2014; Kreeft et al., 2009) and chewing (Epstein et al., 1999). Furthermore, oral cancer treatment can lead to reduced tongue mobility (Kappert et al., 2019), and impaired speech intelligibility (Lazarus et al., 2014; van der Molen et al., 2012). The oral cancer survivors who have impaired speech have more difficulty communicating with other people, which may negatively affect their quality of life (Epstein et al., 1999). Such communication problems could be alleviated with ASR systems, as they make it easier for oral cancer patients to communicate with others despite of their speech impairment.

Although the research interest on ASR performance for oral cancer patients is increasing, oral cancer speech data is difficult to collect and, therefore, studies on the subject are still limited. However, recently a new oral cancer database has been collected which enables assessing ASR performance on oral cancer speech. The current study aims to fill the gap in the literature on ASR performance on oral cancer speech in Dutch by comparing the ASR performance on oral cancer speech and healthy speech. In order to be able to answer our research questions, we conducted extensive error analyses on the phoneme and articulatory feature level in addition to the word error rate. This serves multiple goals:

- First, comparing the results of this analysis for both the healthy and oral cancer speech can reveal what phonemes are difficult to capture for ASR systems in both healthy and oral cancer speech, and what phonemes are hard to capture in oral cancer speech only.
- Second, in order to develop ASR systems that are trained on oral cancer speech, it is essential to know more about the kind of errors that are made, and specifically what types of phonemes in oral cancer speech cause difficulties for ASR systems that are trained on healthy speech. In addition, it is important to investigate whether these sounds correspond to the phonemes that are mentioned in the literature.
- Third, it would be beneficial for the future development of ASR systems for oral cancer speech to look into what phonemes are misrecognized for patients with different types of surgical treatments. This could contribute to either a more inclusive ASR system for oral cancer patients, or to ASR systems that are developed for an even more specific target population.

The outcomes of this study contribute to the development of ASR systems specifically for oral cancer speech in Dutch, which could improve the quality of life of oral cancer survivors.

1.1 Research questions and hypotheses

The present study aims to gain insights into what type of articulatory aspect(s) of oral cancer speech are particularly difficult for standard ASR systems to recognize compared to healthy speech in Dutch. In addition, we aim to investigate the influence of the surgical treatment on the ASR performance on oral cancer speech. This study thus seeks to answer the following research questions:

RQ1. 'What phonemes in oral cancer speech cause higher recognition error rates in a standard ASR system compared to healthy speech in Dutch?'

RQ2. 'Does the surgical treatment of oral cancer patients influence the ASR performance on oral cancer speech?'

Following previous research (e.g. Borggreven et al., 2005; Laaksonen et al., 2011; Halpern et al., 2020, 2022), it is hypothesized that the answer to the first research question is that plosives (mainly /k/, /p/, /t/, and /d/) and alveolar sibilants cause higher recognition error rates in oral cancer speech than in healthy speech in Dutch. In addition, it is expected that certain vowels, especially /a/ and /u/, cause difficulties for the ASR system as well (Halpern et al., 2022).

Furthermore, regarding the second research question it is expected that the type of surgery impacts the ASR performance, as previous research has suggested that the speech intelligibility of oral cancer patients is influenced by the site of resection (Logemann et al., 1993; Borggreven et al., 2005). In addition, a mandibulectomy affects more articulators than a glossectomy (Matsui et al., 2007), which leads us to hypothesize that the ASR performance for these patients is worse than for patients who underwent a (partial) glossectomy. Regarding specific phonemes, research has shown that the jaw is important for the production of vowels (Mooshammer et al., 2007), and we therefore believe that vowels are more impaired in the speech of oral cancer patients with mandibular surgery, resulting in higher recognition error rates.

1.2 Thesis Outline

This thesis contains six chapters. Chapter 1 has given a short introduction to the research topic and the aims and relevance of this research. The second chapter presents an overview of the existing theory in this field in order to create a context and to acquire a better understanding of the results. In Chapter 3, the methods used in this study are described, followed by an overview of the findings in Chapter 4. Then Chapter 5 provides a discussion of the results, which also acknowledges the limitations of this research. Lastly, the sixth and final chapter summarises the findings and draws conclusions.

2 Background Literature

This chapter discusses the literature and concepts related to this study. It is divided into five main topics: Failure of speech production (Section 2.1), Oral cancer speech (Section 2.2), Deep Learning (Section 2.3), Automatic Speech Recognition (ASR) (Section 2.4), and finally, this chapter ends with a review of previous research on ASR for pathological speech (Section 2.5).

2.1 Failure of speech production

When we are talking about pathological speech, we mean speech that is impaired due to a malfunction in the human speech production system. In order to understand how speech production can fail, it is important to have an understanding of how speech is generally produced. There are several theories on how speech is produced, but one that is widely accepted is the source-filter theory (e.g. Fant, 1981). This theory states that the vocal folds in the larynx generate an acoustic source, which is the acoustic energy that serves as the input for the speech production system. This acoustic source is then modulated ('filtered') by the flexible articulators in the vocal tract called 'filters', resulting in the output sound. The flexible articulators are constantly moving, and the specific shape and configuration of the articulators determine the formants, or resonant frequencies of the vocal tract (Seikel et al., 2019). Formants can be defined as the frequencies of a sound that resonate the most in the oral cavity given the shape and position of the articulators, and these formants in turn determine the output sound. Figure 1 gives a simple visualisation of the source-filter theory we have briefly described.

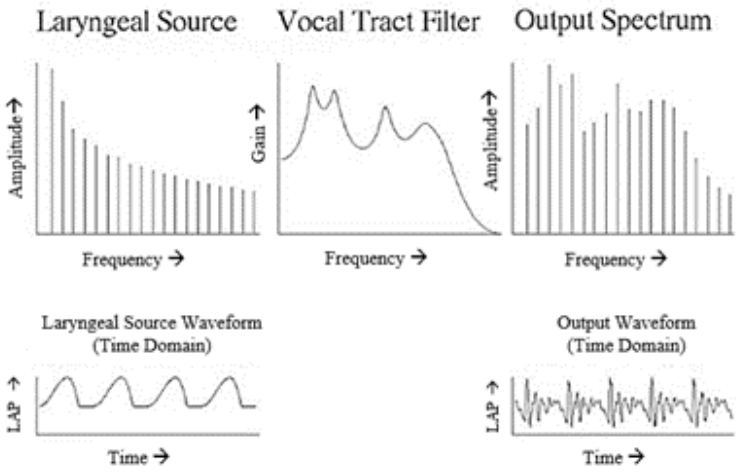


Figure 1: A simple visualisation of the source-filter theory of speech production.

The human speech production system can fail in various ways involving either the source, the filter or both, and this leads to several different speech pathologies. For instance, in pathologies such as dysphonia (e.g. Bender et al., 2004) and dysarthria (e.g. Enderby, 2013) there are issues with voicing or phonation, which means that the source is involved. It must be mentioned however, that in the case of dysarthria these issues are the result of neuromuscular damage (Enderby, 2013). Dysarthria can therefore also affect the articulation, meaning the issues involve the filter of the speech production system as well. Other pathologies that (mainly) affect the filter are for example cleft lip (e.g. Safaiean et al., 2017), cleft palate (e.g. Rullo et al., 2009), and oral cancer speech (e.g. van der Molen et al., 2008), although the latter can also be affected on the phonation level. In addition, besides dysarthria there are other pathologies with a neurological origin that affect aspects of speech, such as Alzheimer (e.g. Lindsay et al., 2021) and aphasia (e.g. Qin et al., 2020). For the purpose of this research, however, we focus on the speech of oral cancer patients who have undergone surgical treatment. This treatment has inflicted trauma to their articulators, and thus causes articulatory difficulties for the oral cancer

patients, due to tissue loss. The next section elaborates on the characteristics of the speech of oral cancer patients.

2.2 Oral cancer speech

Multiple studies have found that treatment for oral cancer can negatively affect speech intelligibility of oral cancer patients (e.g. Lazarus et al., 2014; van der Molen et al., 2012). This is mainly the case when patients have undergone a total or partial glossectomy, in which the entire tongue or parts of the tongue are removed. In addition, research has also shown that oral cancer patients who have had a mandibulectomy, in which a part of the mandible is removed, suffered from significantly impacted speech (e.g. Logemann et al., 1993; Matsui et al., 2007). Sufficient control of the articulators is critical for the production of speech. When any of the articulators are impaired in terms of movement, strength, or adaptability this impacts a speaker's ability to make the articulatory movements that are required to produce speech (Saravanan et al., 2016). In order to limit the extent to which speech organs are affected, different methods of oral cancer treatment have arisen over the past couple of decades, such as reconstruction of the damaged tissue (e.g. free flaps) and organ preservation (e.g. chemoradiation). Nevertheless, in a literature review van der Molen et al. (2008) found that both oral cancer treatments, i.e., reconstruction and organ preservation, often cause speech impairments. More recent approaches to treat oral cancer are radiation delivery techniques and speech rehabilitation, in which the articulatory organs are spared (de Bruijn et al., 2009).

In previous studies, it has been suggested that the amount of removed tissue (e.g. Rentschler and Mann, 1980), the site of the resection (e.g. Logemann et al., 1993), and the technique of reconstruction (e.g. Konstantinović and Dimić, 1998) might be a strong indicator of the resulting speech intelligibility of patients. More recently, Borggreven et al. (2005) also found that the size and location of the tumour appear to highly influence the quality of speech after treatment. For example, patients who were treated for larger tumours experienced more difficulty with their speech compared to patients who were treated for smaller tumours. Furthermore, even though speech is mainly impaired on the articulatory level, Lazarus et al. (2014) found that patients who also underwent radiation therapy experienced issues with phonation as well. In addition, a number of studies found that both swallowing and speech functions worsened over time in patients who underwent both surgery and radiotherapy (e.g. Shin et al., 2012; Lazarus et al., 2013).

There are several characteristics of speech impairment that previous studies have found to be a consequence of oral cancer treatment. The common findings are that primarily plosives (e.g. Bressmann et al., 2004, 2009; de Bruijn et al., 2009) and alveolar sibilants (e.g. Logemann et al., 1993; Laaksonen et al., 2011; Halpern et al., 2020) are affected. Additionally, certain vowels such as the /i/, have been found to be affected as well as the vowel space area (e.g. Whitehill et al., 2006; Takatsu et al., 2017). The following sections further elaborate on each of these affected speech sounds.

2.2.1 Plosives

In a study with German glossectomy patients, Bressmann et al. (2004) found a moderate but significant correlation between tongue motility and consonant intelligibility, which, according to the authors, supports the assumption that better tongue motility indicates better articulation after a glossectomy. Furthermore, the study of Borggreven et al. (2005) with Dutch participants showed that when patients with different kinds of oral cancer produced the velar /k/ sound, this was often perceived as /x/ instead.

de Bruijn et al. (2009) also found that the production of /k/ and /x/ was a good predictor of oral cancer speech in Dutch. They point out that the production of these speech sounds 'require a posterior move of the tongue towards the oropharyngeal region and an adequate motility of the velum' (de Bruijn et al., 2009, p. 184), and that better tongue motility corresponds with better consonant intelligibility (Bressmann et al., 2004). Moreover, in the study of Borggreven et al. (2005) study, it was difficult for oral cancer patients to produce the alveolar /d/ and /t/ as well, as these speech sounds were often nasalized or retracted: /d/ was confused with /n/ (in patients treated for oropharyngeal tumours), and /t/ with /tʃ/ (in patients treated for tongue tumours).

All of the aforementioned studies used human listeners to evaluate the intelligibility of oral cancer speech. However, even though ASR systems could be both faster and cheaper to evaluate the speech intelligibility of oral cancer patients in clinical practices (e.g. Windrich et al., 2008), research on the use of ASR systems to evaluate speech intelligibility of oral cancer patients is still quite limited. Nevertheless, two recent studies have been conducted using ASR systems to evaluate which phonemes are affected most in oral cancer speech. The findings of Halpern et al. (2020) reveal that plosives are among the phonemes that are the most important indicators for oral cancer speech in English. In addition, Halpern et al. (2022) found that when the ASR system is trained on healthy English speech, the /g/ and /p/ have phoneme error rates (PERs) that exceed 60%, and are therefore among the most difficult phonemes to be captured by ASR systems.

2.2.2 (Alveolar) sibilants

Besides the plosives mentioned in the previous section, Borggreven et al. (2005) also found that patients who were treated for tongue tumours had difficulty producing the alveolar sibilant /s/. They tended to retract the consonant and confuse it with /ʃ/. Furthermore, Laaksonen et al. (2011) investigated the effects of reconstructive surgery on the sibilants that were produced by Canadian-English tongue cancer patients. Their results showed that even one year after their surgical treatment, the patients were unable to articulate /s/ and /z/ in a manner similar to their pre-operative speech. However, it is important to keep in mind that these findings concern spectral and temporal acoustic measures, which does not necessarily mean that the intelligibility of these phonemes was affected. Moreover, the analysis of /ʃ/ did not show any significant results and the authors hypothesize that this might be due to the tolerance of /ʃ/ regarding articulatory deviation compared to /s/ and /z/. In addition, they found that shortly after the treatment the acoustic distance between these three sounds was reduced, although this reduction could no longer be observed one year after the reconstruction (Laaksonen et al., 2011).

Besides the plosives, Halpern et al. (2020) found that sibilant frequencies are important indicators for the detection of oral cancer, which is in accordance with previous literature. In contrast with this, however, Halpern et al. (2022) found that /s/ and /z/ were captured relatively well by two ASR architectures trained on healthy English speech. They suggest that due to the fact that sibilants are considered to be noise, the ASR performance is less impacted by loss of information of these phonemes.

2.2.3 Vowels

Whitehill et al. (2006) investigated the acoustic characteristics of vowels of Cantonese oral cancer patients who had undergone a glossectomy. In addition to an intelligibility test conducted with speech and hearing science students, the authors performed an acoustic analysis by measuring the formant

frequencies of the four vowels /i/, /e/, /a/, and /u/. They found that, as a result of reduced tongue movement after a partial glossectomy, the production of the vowel /i/ was the most affected vowel in oral cancer speech. This was due to the fact that it had a significantly reduced F2 value compared to the speech of the control speakers and it was reportedly the least intelligible vowel as well. However, it was not just the production of the /i/ that was affected, as Whitehill et al. (2006) found that the entire vowel space area was compressed following the glossectomy. In accordance with these findings, de Bruijn et al. (2009) found similar results in the case of Dutch oral cancer speech, i.e. compressed vowel space area and affected production of /i/.

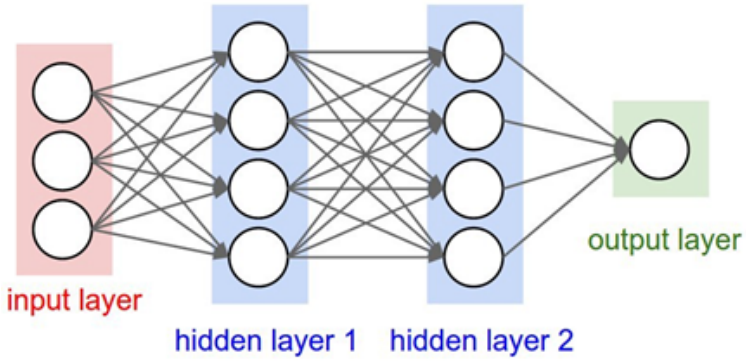
Takatsu et al. (2017) carried out a similar study on vowel production with Japanese speakers. In comparison with the study of Whitehill et al. (2006) and de Bruijn et al. (2009), (some) speakers had also received reconstructive surgery in Takatsu's study. Nevertheless, Takatsu et al. (2017) found that the vowel space area and the diphthongs were still influenced by the surgical treatments, albeit to different extents. However, Takatsu et al. (2017) also show that patients who received reconstructive surgery had a larger vowel space area compared to patients who did not.

In contrast to these studies, the findings of Halpern et al. (2020) do not support the significance of vowels and diphthongs when they performed an oral cancer speech detection task. Although, when Halpern et al. (2022) performed an error analysis on two ASR architectures trained on healthy speech, the analysis did reveal that certain vowels were hard to capture for these systems. The hybrid model had difficulties capturing the /a/ and /u/, and the end-to-end architecture had issues recognizing the /a/, /ɛɪ/ and /u/. Nevertheless, the authors mention that, apart from the /a/ and /u/, the vowels were recognized comparatively well.

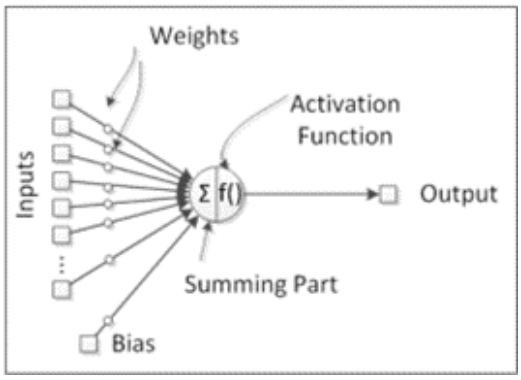
2.3 Deep Learning (DL)

The field of deep learning concerns the development of machine learning (ML) techniques using mutli-layered neural networks for various tasks (Alzubaidi et al., 2021). Deep learning algorithms are inspired by the way human brains process information, and the neural networks involved in deep learning thus have a similar structure (Sugomori et al., 2017). Therefore, as the human brain is able to perform tasks based on learned knowledge, in order to find the relationship between the input and output data of a system, rather large amounts of input data are required for the deep learning algorithm to learn from (Alzubaidi et al., 2021). Figure 2a is a simplified representation of a multilayer perceptron which is the simplest deep neural network (DNN), although state-of-the-art neural networks (e.g. transformers) typically consist of more than two hidden layers as well as various other complex components (e.g. Zhang et al., 2020).

In simple terms, the input layer stores the input, which is processed by the hidden layers and then the output layer provides the desired output. In more technical terms, neural networks aim to learn a function that maps the input to the output. The function that is being learnt by the neural network is determined by the manner in which the network is structured as well as by the weights inside the network, which can either amplify or attenuate the inputs (Sugomori et al., 2017; Zhang, 2022). Furthermore, neural networks have a loss function that indicates how closely the learnt mapping function approximates the optimal mapping. Therefore, the training of a neural network can be described as an optimization process, which means that the network weights are repeatedly updated in order to minimize the loss function (Alzubaidi et al., 2021; Zhang, 2022).



(a) A simple visualization of a neural network.



(b) *A visualization of a hidden unit.* (Note. Reprinted from 'Deep Learning' by Sugomori et al., 2017, p. 466)

Figure 2: A simplified representation of how a neural network is structured.

Neural networks thus contain hidden layers and these hidden layers in turn consist of small hidden units. Each hidden unit calculates a weighted sum of the previous layer (see Figure 2b; Sugomori et al. (2017)). The weights of this weighted are learnt during an optimization process. Then, the hidden unit mathematically adds non-linearity by using an activation function that fires the output (Sugomori et al., 2017). There are many different activation functions, though the most commonly used activation functions are Sigmoid, Tanh (Hyperbolic Tangent) and ReLU (Rectified Linear Unit) for hidden layers, and Linear or Softmax for the output layer. The activation function fires the output based on the weighted sum and the bias, which is an independent parameter that acts like input, although it is stimulated by a fixed value that is multiplied by an associated weight (Sugomori et al., 2017). Finally, the output of every hidden unit in the hidden layer serves as input for the following neural network layer.

Deep learning can be applied in many different situations, and it sometimes even outperforms human experts (Alzubaidi et al., 2021). Some examples of the use of deep learning across industries are self-driving cars, biometrics, fraud detection, and of course automatic speech recognition. Although ASR is usually associated with voice assistants, it has a much wider application. For example, ASR systems can be employed in clinical practices to estimate speech intelligibility for example, and these tests would be both cheaper and faster than using human listeners. The next section dives deeper into automatic speech recognition and the role deep learning plays in it.

2.4 Automatic Speech Recognition (ASR)

Automatic speech recognition (ASR) can simply be defined as the process of converting speech into text, i.e. Speech-to-Text (STT), where the ASR system processes human speech as input, recognizes it, and then gives the corresponding word sequence as output. In other words, an ASR system aligns speech and text by identifying and classifying the input data, which means that we can consider ASR to be a classification task. Thus, given the speech signal $\mathbf{X}$ and its transcriptions $\mathbf{Y} = (y_1, y_2, \ldots, y_L)$ (where the elements are the words within the text sequence), ASR systems learn how to model the conditional probability distribution $P(\mathbf{Y}|\mathbf{X})$ (Jurafsky and Martin, 2020). The ASR system obtains the predicted transcription $\mathbf{Y}^*$ with the following formula:

$$\mathbf{Y}^* = \arg\max_{Y} P(\mathbf{Y}|\mathbf{X}^*)$$

There are two main approaches to ASR, namely the traditional approach and the end-to-end deep learning approach. The traditional approach dominated the field of ASR before deep learning was introduced. Figure 3 gives a simplified representation of the pipeline of a traditional ASR system, which consists of separate components. First, acoustic features are extracted from the speech signal, which thus contain information of the speech signal within a specific time frame (Zhang, 2022). These feature vectors serve as input for the decoder, which aims to efficiently find the optimal text sequence given the feature vectors. The decoder consists of the acoustic model, a pronunciation dictionary or lexicon, and a language model. First, the acoustic model models the likelihood of a sequence of sound units based on the given feature vectors. Then, the pronunciation dictionary maps the given sound units into text, and the language model assigns the probability of this text sequence occurring in the language. The decoder then outputs the optimal text sequence matching the input speech.

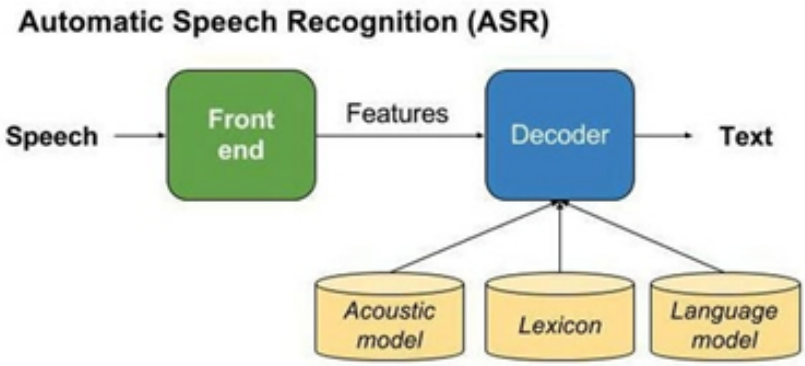


Figure 3: A simple visualization of a traditional ASR pipeline.

The introduction of deep learning has significantly benefited the field of ASR (Graves and Jaitly, 2014). Initially, deep neural networks were only used for acoustic modelling. ASR developers believed that simplifying the different components of the ASR pipeline into a single neural network could improve the ASR performance (Watanabe et al., 2017). These simplified systems are called end-to-end ASR, and the next section discusses these end-to-end ASR systems in more detail.

2.4.1 End-to-end ASR

The goal of an end-to-end (E2E) ASR system is to directly map sequences of feature vectors into word or sub-word sequences, rather than using several modules to achieve this (Watanabe et al., 2017). In order to deal with vocabulary related issues and to improve generalization, E2E ASR models generally use character representations for the output sequence instead of word representations (Watanabe et al., 2017). Although a disadvantage of E2E ASR systems is that, in contrast to traditional ASR systems, they require large amounts of data (Watanabe et al., 2017). Nevertheless, the simplicity and high recognition accuracy of E2E ASR systems has increased their popularity over the past couple of years (Deng et al., 2022). The current state-of-the-art ASR is a variant of a Transformer architecture, called Conformer (see Gulati et al. (2020)).

Within the E2E ASR approach, a commonly used loss function is the Connectionist Temporal Classification (CTC). Since neural networks in ASR are typically trained to classify the frame-level feature vectors (see Section 2.4), this indirectly requires an alignment between the audio and the transcription sequences (Graves and Jaitly, 2014). However, this alignment would only be reliable after the classifier, i.e. the neural network, is trained (Graves and Jaitly, 2014). This is where CTC comes into play, as the dynamic programming in CTC allows for an efficient calculation of both the log probability and its gradient, which can be propagated for learning recurrent neural network (RNN) parameters (Watanabe et al., 2017). In other words, CTC is a loss function that allows the training of neural networks for sequence-level transcription tasks without the requirement of prior alignment of the input and output sequences (Graves and Jaitly, 2014).

CTC uses an output layer that has the intermediate label representation that allows repetitions of transcription labels (e.g. characters, phonemes, words) as well as the occurrence of a 'blank' ('_'), which is a special emission without labels (Graves and Jaitly, 2014). Given an input sequence $\mathbf{X} = (x_1, x_2, \ldots, x_T)$, the probability $P(\pi|\mathbf{X})$ of a CTC alignment $\pi = (\pi_1, \pi_2, \ldots, \pi_T)$ where the elements represent the intermediate label representations, is the product of the emission probabilities at every time step (Graves and Jaitly, 2014):

$$P(\pi|\mathbf{X}) = \prod_{t=1}^{T} P(\pi_t, t|\mathbf{X})$$

The 'blank' in CTC explicitly represents the boundary of a transcription label in order to deal with the repetition of these labels (Watanabe et al., 2017). Moreover, when a label is repeated in two or more successive time frames in alignments, the repetitions of the label are deleted (Graves and Jaitly, 2014). For instance, 'hhheeeeeell' could be decoded as either 'hell' or 'heel' in CTC. However, with the introduction of the 'blank' we can separate the two transcriptions with '_hheeeel_lll' mapping to 'hell', while 'hhhee_eeeellll' maps to 'heel'. Thus, the introduction of 'blank' labels additionally allows us to distinguish between different time-alignments, i.e. the short and long 'e' in our example. We can formulate the probability $P(\mathbf{Y}|\mathbf{X})$ of an output transcription $\mathbf{Y} = (y_1, y_2, \ldots, y_L)$ when we use an operator Φ that removes the repeated labels and then the blanks:

$$P(\mathbf{Y}|\mathbf{X}) = \sum_{\pi \in \Phi^{-1}(\mathbf{Y})} P(\pi|\mathbf{X})$$

We can observe that the probability of Y equals the sum of probabilities of the corresponding alignments. The idea behind this is that since it is unknown where the labels will occur in a particular transcription, we sum over every possible place of occurrence (Graves and Jaitly, 2014). Using dynamic programming, we can then train a network to minimize the CTC loss function for a given target transcription $\mathbf{Y}^*$:

$$CTC(\mathbf{X}) = -\log P(\mathbf{Y}^*|\mathbf{X})$$

Attention-based E2E ASR In contrast to other deep neural networks that deal with fixed-length inputs and outputs, recurrent neural networks (RNNs) are structured to exploit important contextual information and are therefore powerful models to process sequential data (Graves et al., 2013). This is useful in dealing with speech data, since knowing previously uttered words can be helpful for recognizing currently spoken words. RNNs operate with a hidden state vector, which means that the hidden layers of RNNs do not only forward their output to the following neural network layer, but back to

the hidden layer itself as well (see Figure 4). Loops like this allow the hidden layers to iteratively store contextual information through previous temporal steps, and the hidden layers can therefore be considered to be short-term memory units (Zhang, 2022).

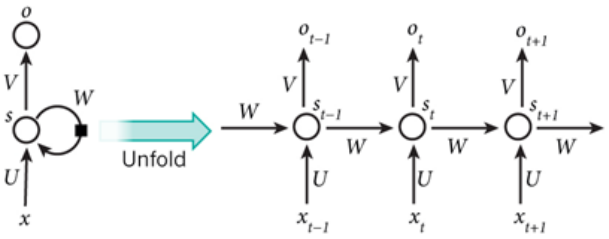


Figure 4: A visualization of a recurrent neural network (RNN).

However, RNNs deal with the problem of vanishing or exploding gradients (e.g. Chung et al., 2014). This issue was alleviated with long short-term memories (LSTMs; Hochreiter and Schmidhuber (1997)) and gated recurrent units (GRUs; Cho et al. (2014)) by implementing so called 'gating mechanisms' that bypass many temporal steps and control the flow of error (Chung et al., 2014). This was mainly inspired by neurological processes, although it also had some mathematical advantage (Chung et al., 2014). Nevertheless, these solutions still required alignments. This problem was in turn alleviated by bidirectional LSTMs (BLSTMs; e.g. Graves et al. (2013)) and light GRUs (Li-GRUs; Ravanelli et al. (2018)), which are non-causal counterparts of LSTMs. This means that, for example, the LSTM could not utilize its fifth input to make decisions at output 1 as it was not able to 'see' it, whereas BLSTM is able to access long-range context in both directions (Graves et al., 2013). Unfortunately, even though BLSTMs and Li-GRUs did not require alignments, their time dependency was non-ideal. This is when attention-based architectures were introduced.

When an ASR architecture is attention-based, this means that an attention mechanism is implemented to align the acoustic frames and the recognized labels, and there are no conditional independence assumptions required (Watanabe et al., 2017). In simple terms, attention mechanisms 'learn to focus their "attention" to specific parts of their input' (Bahdanau et al., 2016, p. 4945). At every time step i, attention mechanisms generate a context vector c_i, which captures the information in the acoustic signal that is required to generate the following character (Chan et al., 2016). This allows the attention mechanism to achieve larger time dependencies, which is essential as contextual information is crucial to perform numerous tasks such as speech recognition and translation (e.g. Graves et al., 2013). Additionally, the label synchronous prediction is what characterizes attention mechanisms (Miao et al., 2019), and this prediction is derived from 'attending' to segments of the input (see Figure 5).

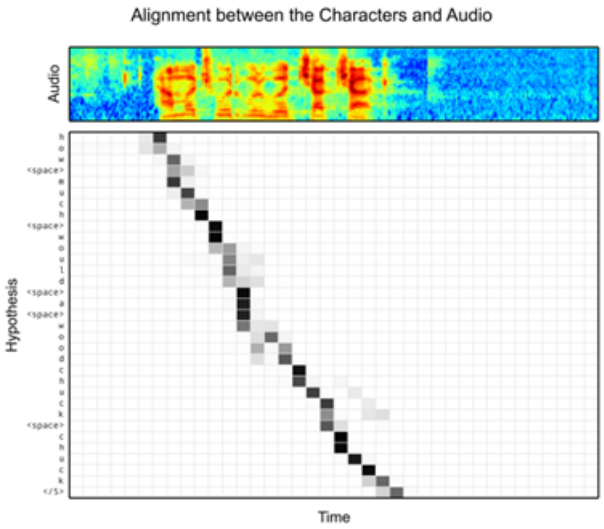


Figure 5: *Alignments between character outputs and audio signal produced by the Listen, Attend and Spell (LAS) model for the utterance "how much would a woodchuck chuck".* (Note. Reprinted from 'Listen, Attend and Spell' by Chen et al., 2016, p. 4963)

Currently, E2E ASR architectures that rely on both attention and CTC are very popular, since these architectures exploit the best properties of both (Deng et al., 2022). In this study, we adopt such an E2E ASR system, namely ESPnet. Figure 6 gives an overview of the standard recipe flow in ESPnet, which is relatively simple due to the benefits of E2E ASR. The six stages that a standard recipe follows are the following (Watanabe et al., 2018):

- **Stage 0 - Data preparation:** the data is prepared by adopting the Kaldi-style data directory format. In addition, the Kaldi data preparation script can be used as well.
- **Stage 1 - Feature extraction:** acoustic features are extracted using Kaldi. The majority of the recipes extract the 80-dimensional log Mel feature and the pitch feature, which results in a total of 83 dimensions.
- **Stage 2 - Data preparation for ESPnet:** in this stage all of the information (including the information in the Kaldi data directory) is converted into a single JSON file, with the exception of the input features.
- **Stage 3 - Language model training:** this is the only optional stage and therefore there are multiple recipes that do not have it. This stage trains the character-based RNNLM with either the Chainer or the PyTorch backend.
- **Stage 4 - E2E ASR training:** the Chainer or PyTorch backend is used to train the hybrid CTC-attention-based encoder-decoder.
- **Stage 5 - Recognition:** in case the RNNLM was obtained in stage 3, it will be used together with the E2E ASR model obtained in stage 4 in order to perform speech recognition.

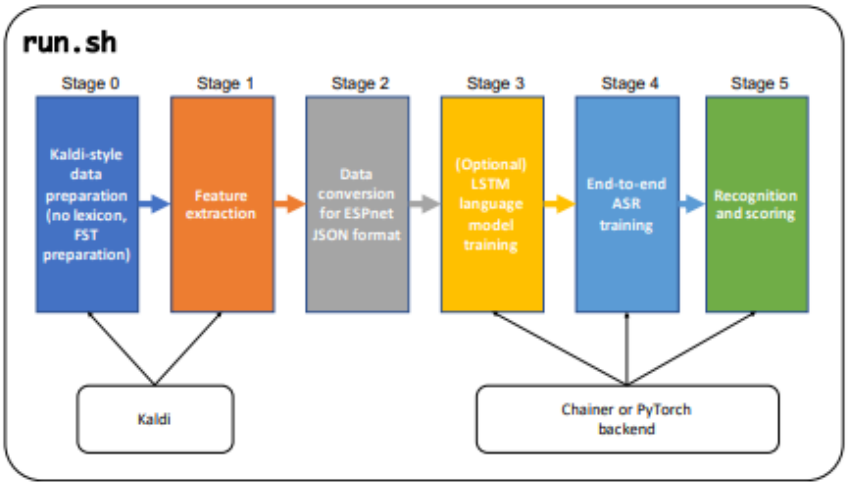


Figure 6: *Experimental flow of standard ESPnet recipe.* (Note. Reprinted from 'ESPnet: End-to-End Speech Processing Toolkit' by Watanabe et al., 2018, p. 2209)

2.5 ASR for pathological speech

Oral cancer speech is not the only type of pathological speech that has sparked interest in the field of automatic speech recognition. Pathological speech in general has challenged researchers for years, as it poses extra difficulties for the development of ASR systems, due to the impaired speech intelligibility and the lack of training data (low-resource). Nevertheless, in recent years there has been an increase in the research and development of ASR systems specifically for pathological speech. Pathological ASR research includes research on dysarthric ASR systems (e.g. Sharma and Hasegawa-Johnson, 2013; Calvo et al., 2020; Hermann and Doss, 2020), as well as research on ASR systems developed for aphasia (e.g. Qin et al., 2018), and oral cancer speech (e.g. Maier et al., 2010).

For dysarthric speech, Christensen et al. (2013) found that by using healthy speech instead of dysarthric speech to train the feature-generating neural network they were able to improve the acoustic modelling of their ASR system. In addition, the results of Hermann and Doss (2020) revealed that after training their acoustic models using the lattice-free loss function, the recognition performance of their ASR system had improved compared to conventional training methods. Furthermore, Yilmaz et al. (2017) employed a multi-stage DNN training in order to develop acoustic models for pathological speech that are more robust, in which they initially trained the model on more general data (i.e. healthy speech) and then retrained it on dysarthric speech to develop a domain-specific model. They found that their multi-stage training approach obtained a better recognition performance than their baseline systems that were trained on either healthy or dysarthric speech. For aphasic speech, Qin et al. (2018) found improvement in performance as well when they adopted a TDNN-BLSTM architecture for acoustic modelling and employed a multi-task learning technique in which they used a large amount of healthy speech data.

It must be kept in mind, however, that most of the aforementioned studies investigated hybrid ASR models rather than E2E ASR systems. Even though there are some studies on E2E ASR systems for pathological speech (e.g. Harvill et al., 2021; Qin et al., 2020), the amount of literature on this topic is still relatively limited. Moreover, the transformer-based ASR system used by Harvill et al. (2021) even yielded higher WERs than the state-of-the-art ASR described by Hermann and Doss (2020). Although, in his recent study Shahamiri (2021) found that when they converted the word utterances into visual feature representations and then attempted to recognize the visual representations rather than phonemes, this improved the recognition accuracy of an E2E ASR system for mild and severe dysarthric speech. In addition to the conversion to visual representations, Shahamiri (2021) applied transfer learning by using healthy speech to learn the visual representations for the dysarthric speech.

2.5.1 ASR for oral cancer speech

As mentioned above, the amount of research on ASR systems for different kinds of pathological speech has increased in the past couple of decades, including oral cancer speech. For example, Windrich et al. (2008) investigated the word recognition rates of an ASR system for oral cancer speech and healthy speech. They compared the ASR performance to the perceptual evaluation of experts in order to find out whether an ASR system would be a valuable tool to objectively and quantitatively evaluate the speech of patients after they have received treatment for oral cancer, which would serve both clinical and research purposes. Their results revealed that the ASR system recognized between 8% and 82% of the oral cancer speech, and that the performance of the ASR system correlated closely with the perceptual intelligibility evaluation of the experts (r=-0.93). Similarly, Maier et al. (2010) investigated how applicable ASR systems were to speech disorders that were caused by head and neck

cancers. Their findings confirmed those of Windrich et al. (2008), as they found that the speech of the oral cancer patients yielded significantly higher WERs than the speech of the healthy controls, and that these results correlated highly with the evaluation of experts.

More recently, Halpern et al. (2022) investigated two ASR architectures and the effect of three different AM approaches on an oral cancer ASR task. Their baseline systems consisted of one hybrid deep neural network-hidden Markov model (DNN-HMM) and one E2E ASR model that follows a transformer architecture, and both of these systems were trained on healthy English speech only. The approaches that the authors investigated were a retraining approach for both architectures, and a speaker adaptation and disentangled representation learning approach for the hybrid architecture only. Their results revealed that the speaker adaptation approach outperformed the other approaches and achieved the biggest improvement compared to the baseline system trained on healthy speech, followed by the retraining approach using the E2E architecture. The speaker adaptation approach achieved a WER of 62.8% on oral cancer speech, which is a 7.8% absolute word error rate (WER) reduction in comparison to the baseline system. The E2E retraining approach obtained a WER of 63% and this was a 7.5% absolute improvement over the baseline E2E ASR.

3 Methodology

In this chapter, the dataset (Section 3.1) as well as the model (Section 3.2) used in this study is described, followed by an overview of the error analyses that were performed on the phoneme and articulatory feature level (Section 3.3).

3.1 Dataset

For this research, two datasets were used: the NKI-UMCG-RUG corpus containing Dutch oral cancer speech as well as healthy speech, and the CGN corpus, consisting of typical Dutch speech. The latter was used as training data for the ASR system and the former as test data.

3.1.1 Oral cancer speech dataset

We used the NKI-UMCG-RUG oral cancer speech corpus, which was created within the project '*Articulation and coordination of speech after treatment for oral cancer*'. This project is a collaboration between the Netherlands Cancer Institute (Dutch: *Nederlands Kanker Instituut* (NKI)), the University Medical Center Groningen (UMCG), and the University of Groningen (Dutch: *Rijksuniversiteit Groningen* (RUG)). In order to ensure that the data was treated with the utmost care by the researcher and the privacy of the participants would thus not be violated, a data agreement was signed (see Appendix A).

The dataset contains eleven Dutch speakers: six speakers who have been treated for oral cancer (i.e. post-treatment) and five control speakers. All of the speakers come from northern regions of the Netherlands. Table 1 gives an overview of the age and gender of the participants per condition and in total.

Table 1: Overview of the participant characteristics per condition and overall.

	Healthy (n = 5)	**Patient (n = 6)**	**Overall (n = 11)**
Age			
Mean	61.6	59.5	60.45
Range	56-77	47-75	47-77
Gender			
Female	3 (60%)	3 (50%)	6 (54.55%)
Male	2 (40%)	3 (50%)	5 (45.45%)

Within the patient group, three of the participants had undergone a mandibulectomy, and the other three had undergone a glossectomy. Of those three patients who had tongue surgery, only one had a tongue reconstruction. Moreover, all of the patients except for one had received either chemotherapy or post-operative radiation therapy.

The recordings contain approximately 30 minutes of speech for every speaker, consisting of 207 sentences that were read out loud. There were three types of sentences: literary sentences, news sentences, and masking noise sentences. The literary sentences were taken from five different stories

of varying lengths and the news sentences were retrieved from several Dutch news articles (see Appendix B). The masking noise sentences (i.e. MASK1, MASK2, MASK3, MASK4N, MASK5N, and MASK6N), however, were not included in this study, since these sentences were relatively unnatural compared to the other types of sentences. Moreover, the latter half was produced under masking noise, and this data therefore differs acoustically from the rest of the data. In addition, in case a sentences was rerecorded, the original audio file was omitted from the data as well, and the rerecorded file was used instead. Therefore, a total of 202 utterances per participant was used, resulting in a final dataset of 2222 utterances.

Data pre-processing Before running the data through ESPnet, the oral cancer speech data was pre-processed[1] using librosa Python library (McFee et al., 2015). The original corpus has a sampling rate of 44.1 kHz, but since the model was trained on data with a sampling of 16 kHz, the dataset was resampled to 16 kHz as well. Then the wav-files were mixed from stereo to mono, by taking the average of the two channels. Following this, the loudness of the data was also normalized using librosa. This pre-processed data then served as input data for the ASR system of ESPnet.

3.1.2 CGN corpus

The CGN corpus (Dutch: *Corpus Gesproken Nederlands* (CGN)) is a dataset that contains non-pathological spoken standard Dutch from the Netherlands and Flanders (Nederlandse Taalunie, 2004). The corpus consists of approximately 900 hours of both read and spontaneous speech, of which around one third is data from Flanders and two thirds from the Netherlands. For this study, only speakers from the Netherlands were included, resulting in around 600 hours of speech data.

3.2 Model

In this study, we ran the oral cancer speech dataset through a standard end-to-end ASR system that was pre-trained on the CGN corpus. Then extensive error analyses were performed to gain more insights into the type of errors that are made by a standard E2E ASR system for Dutch oral cancer speech.

3.2.1 End-to-end ASR: ESPnet

We used the ESPnet ASR system, which adopts a hybrid CTC-attention E2E ASR architecture (Watanabe et al., 2017; Miao et al., 2019), and by doing this it fully benefits from the advantages of both implementations in training as well as decoding. In the training process a multi-objective learning framework is used in order to obtain irregular alignments that are more robust and to reach fast convergence (Watanabe et al., 2018). Furthermore, during decoding both the attention-based and the CTC scores are combined in a one-pass beam search algorithm with the intention of further eliminating the occurrence of irregular alignments (Watanabe et al., 2018). In addition, ESPnet uses the dynamic neural network toolkits Chainer (Tokui et al., 2015) and Pytorch (Paszke et al., 2017) as its main deep learning engine, and for the processing of data, the feature extraction and the recipes, ESPnet adopts the style of the Kaldi toolkit (Povey et al., 2011).

In combination with ESPnet, we used a model that was pre-trained on the CGN dataset only. This pre-trained model is a variant of the Transformer architecture, called Conformer (Gulati et al., 2020). The

[1] All the code written for the purpose of this study will be made available here under the name '*DOC-error_analyses*'.

conformer model parameters are as follows: 12 encoder layers, and 6 decoder layers, all with 2048 units. Furthermore, the attention dimension is 256 and there are 4 attention heads. The conformer architecture has a convolutional module with 15 kernels, and has 'two macaron-like feed-forward layers with half-step residual connections sandwiching the multi-headed self-attention and convolution modules' (Gulati et al., 2020, p. 1). The model is trained with 20 epochs. Lastly, we created Kaldi-style recipes for the oral cancer speech dataset in order to be able to run the data through the model.

3.3 ASR evaluation: error analyses

In order to evaluate the ASR performance, error analyses were conducted. First, the word error rate (WER) of the oral cancer speech was compared to the WER of the healthy speech, and then extensive error analyses on the phoneme and articulatory feature level were performed. As we previously mentioned in the introduction, this was done in order to find out (a) what phonemes are difficult to capture for E2E ASR systems in both healthy and oral cancer speech, (b) what types of phonemes are challenging for standard ASR system in oral cancer speech specifically, and lastly (c) what influence the surgical treatment has on the recognition errors that are made for oral cancer speech. Thus, besides the WER, the recognition performance of the ASR system was measured by using the phoneme error rate (PER) and articulatory feature error rate (AFER) as well.

For the first goal, we compared the results of the healthy speech with the results of the oral cancer speech and reflected on the similarities between the two types of speech. The outcomes of the analysis provide important information on which phonemes are hard to recognize in both healthy and oral cancer speech. For the second goal, we looked into the differences between healthy and oral cancer speech, and we therefore focused on the recognition errors made for the oral cancer speech and see whether our findings correspond to the existing literature. For the third and last goal, the ASR performance was compared for the speech of the patients who underwent a glossectomy and the speech of patients who had undergone a mandibulectomy. The analysis reveals which (types of) phonemes were consistently mis-recognized for each patient group.

3.3.1 Word error rate (WER)

The word error rate (WER) is one of the most widely used automatic evaluation measures to evaluate the performance of an ASR system. It is defined as follows:

$$WER = \frac{S+I+D}{N},$$

Where S stands for substitutions, I for insertions, D for deletions, and where N is the total number of words in the reference sentence. The alignment between the reference and hypothesised sentences is done using the Levenshtein distance, also called the Levenshtein alignments. The Levenshtein alignments output the number of insertions, deletions and substitutions that have to take place in order to obtain the hypothesised sentence from the reference sentence (Berger et al., 2021). The WERs were automatically calculated by Kaldi.

3.3.2 Phoneme error analysis and articulatory feature error analysis

Following Halpern et al. (2022), we conducted a phoneme and articulatory feature error analysis in this study. In order to be able to obtain the PERs and AFERs, we performed a grapheme to phoneme conversion, using the phonemizer developed by Bernard and Titeux (2021). Then we aligned the reference and hypothesised sentences with SCTK's program sclite. Once we had these alignments, we were able to calculate the phoneme error rates (PERs) as described in Halpern et al. (2022). The most common definition of the PER is very similar to that of the WER, as it is defined as the sum of insertions, substitutions and deletions divided by the total number of phonemes in the reference sentence:

$$PER = \frac{insertion + substitution + deletion}{N}$$

The PER was calculated for every individual phoneme in order to gain insights into what specific phonemes are difficult to capture for the ASR system.

The articulatory feature error rate (AFER) was calculated in a similar manner as the PER and WER, except that we converted the aligned phoneme sequences to feature sequences based on place of articulation (PoA) and manner of articulation (MoA) before we calculated the error rates (Halpern et al., 2022). The PoA and MoA feature sequences can be found in Table 2. Additionally, the AFERs were calculated for every individual articulatory feature as well, e.g. for the fricatives:

$$AFER_{fricatives} = \frac{insertion_{fricatives} + substitution_{fricatives} + deletion_{fricatives}}{N_{fricatives}}$$

In this study, we present the means and standard deviations of both the PER and AFER.

Table 2: Overview of the phonemes in Dutch. Abbreviations (left to right): Bilabial, Labiodental, Alveolar, Post-alveolar, Palatal, Velar, Glottal.

MoA	**PoA**						
	B	LD	A	P	PAL	V	G
Plosive	p b		t d			k ɡ	ʔ
Nasal	m		n			ŋ	
Trill			r				
Fricative		f v	s z	ʃ ʒ		x ɣ	h
Affricate				tʃ dʒ			
Approximant	w	ʋ	l		j		

4 Results

This chapter presents the WER results (Section 4.1) of the E2E ASR system on the NKI-UMCG-RUG oral cancer speech corpus, as well as the findings of the phoneme error analysis (Section 4.2), and the articulatory feature error analysis (Section 4.3).

4.1 Word Error Rate

The WERs (%) achieved by the baseline E2E ASR system on the oral cancer speech dataset are shown in Table 3. For every participant the percentage of correctly recognized words is listed as well as the percentage of substitutions, deletions, insertions and the WERs. In addition, the average WER results are given for the two speaker groups, i.e. healthy and patient, and overall. As expected, the ASR system performed much better on the healthy speech in comparison to the oral cancer speech. For the healthy speakers the WER ranged from 14.9% to 22.4%, and for the oral cancer speakers it ranged from 37.9% to 93.6%, which means that the highest WER for the healthy speakers is lower than the lowest WER of the oral cancer patients. A non-parametric Mann Whitney U test was conducted in order to find out whether the difference between the WERs of the healthy speakers and oral cancer patients was significant. The statistical test revealed that the average WER of healthy patients (M=17.4) was indeed significantly lower (W=0, p=0.0043) than the WER of the oral cancer patients (M=62.3). This shows that there is a serious performance gap of standard ASR systems on healthy and oral cancer speech.

4.1.1 Effect of surgical treatment on the WER

In order to get more detailed insights into the errors that the ASR system made for oral cancer speech and to find out to what extent the type of surgical treatment influenced the ASR performance, we took a closer look at the results of the oral cancer patients. We divided the oral cancer patients into two groups of three, where one group had undergone a mandibulectomy and the other had undergone a (partial) glossectomy.

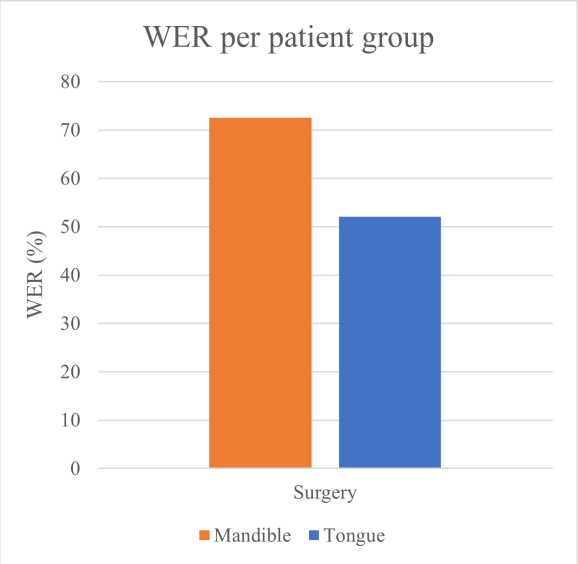


Figure 7: WER (%) results for oral cancer patients grouped by surgical treatment.

Figure 7 shows the WERs of patients who have undergone mandibular surgery and tongue surgery. We observe that the speech of patients who have had tongue surgery was better recognised than that of patients who have had mandibular surgery. For the patients with tongue surgery the ASR system achieved a WER of 52.1% (SD=16.1), while it obtained a WER of 72.5% (SD=30.2) for the patients who underwent a mandibulectomy. Even though these WERs might seem to differ greatly, an independent samples t-test revealed that the difference between the two patient groups failed to reach significance (t (4) = 1.03, p = 0.36), 95% CI [-34.44, 75.18].

Table 3: Overview of the word recognition errors in percentages. **Blue bold** numbers indicate for which participant in each speaker group the ASR system achieved the best performance per column. **Orange bold** numbers represent the worst ASR performance per column for both speaker groups.

	Correct	Substitutions	Deletions	Insertions	WER
Healthy (n = 5)					
01	87.8	10.9	**1.2**	**2.8**	15
07	87.1	11.6	**1.3**	3.7	16.7
08	85.9	12.8	**1.3**	3.9	18
09	**84.7**	**14**	**1.3**	**7.1**	**22.4**
12	**89.5**	**9.3**	**1.2**	4.4	**14.9**
Mean	*87*	*11.7*	*1.3*	*4.4*	*17.4*
SD	*1.8*	*1.8*	*0.1*	*1.6*	*3.1*
Patient (n = 6)					
02	33.8	**62.7**	3.5	**27.4**	**93.6**
03	66.3	30.5	3.3	11.9	45.6
04	44	51.9	4.1	14.4	70.4
05	70.3	27	2.7	**8.2**	**37.9**
06	**72.7**	**26.1**	**1.2**	13	40.3
11	**32.8**	59	**8.2**	18.7	85.9
Mean	*53.3*	*42.9*	*3.8*	*15.6*	*62.3*
SD	*18.6*	*16.9*	*2.4*	*6.7*	*24.3*
Overall (n = 11)					
Mean	*68.6*	*28.7*	*2.7*	*10.5*	*41.9*
SD	*22*	*20.2*	*2.1*	*7.6*	*29.2*

4.2 Phoneme error analysis

In our study, only the phonemes that had a total number of occurrences higher than 100 for both speaker groups, i.e. the healthy speakers and the oral cancer patients, were included in the analysis. This resulted in a total of eight phonemes being omitted from the error analysis. Figure 8 presents the results of the error analysis on the phoneme level, with the y-axis showing the phoneme error rate (PER) and the x-axis indicating the phonemes included in our analysis, grouped by their manner of articulation. The figure shows that for healthy speech, the E2E ASR system achieved PERs between 0% and 10% for the majority of the phonemes. Therefore, we consider phonemes with a PER over 10% to be poorly recognized for healthy speech. Phonemes that have a PER higher than the 10% threshold are /i(ː)/, /y/, /ɛ/, /ŋ/, /h/, and /j/. For oral cancer speech, most phonemes obtained PERs between 25% and 45%, and we therefore set a threshold of 45% in order to classify whether a phoneme was recognized poorly. The phonemes that correspond to PERs over 45% are /i(ː)/, /k/, /ŋ/, and /j/.

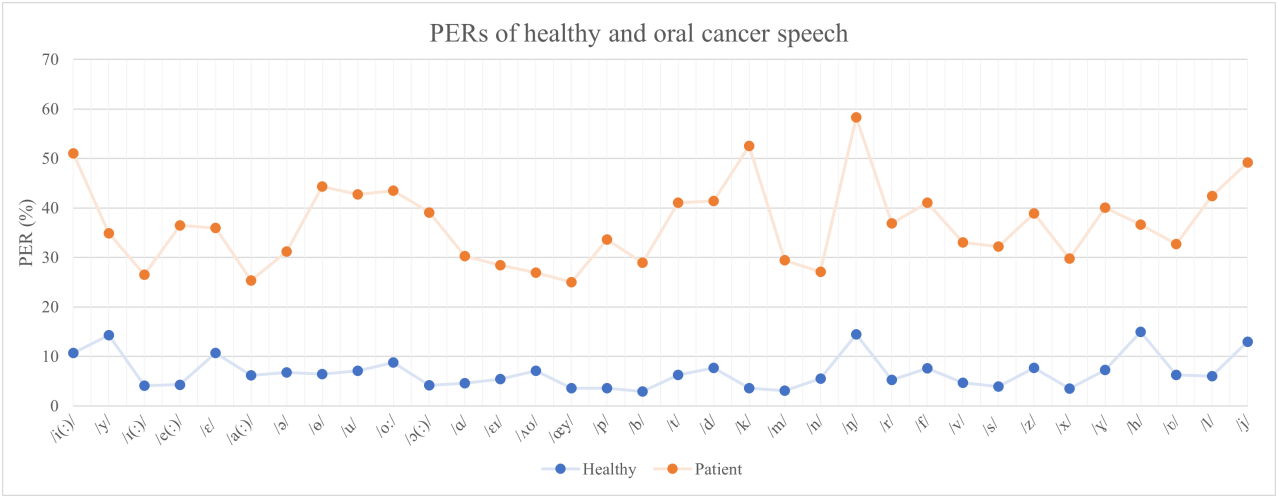


Figure 8: PER (%) results for healthy speakers and oral cancer patients.

4.2.1 Effect of surgical treatment on the PER

For the oral cancer speech data, we have compared the results of the patients who have had tongue surgery with patients who have undergone a mandibulectomy (see Figure 9). It can be observed that, in general, the ASR system achieved better PERs for the patients who had tongue surgery, except for /i(ː)/, /k/, and /ŋ/. These three phonemes were also the only phonemes for the patients with a glossectomy that have a PER over 50%. For the patients with a mandibulectomy there were several phonemes with a PER exceeding 50%, namely /ɵ/, /oː/, /t/, /d/, /ŋ/, /z/, /h/, and /j/. Of these phonemes, the /t/, /d/, and /ŋ/ elicited the highest PERs.

When looking at the PER results per patient (see Appendix C), we can see that the worst PERs were consistently achieved for three patients, namely patient 02, 04, and 11, which is in accordance with the WER findings. It stands out that the velar phonemes of patient 04 were misrecognized most often, while patient 02 had the lowest PERs for almost all of the vowels. For the three patients with the best PERs on the other hand, we can see that the speech of patient 05 obtained the best PERs for all of the velar phonemes.

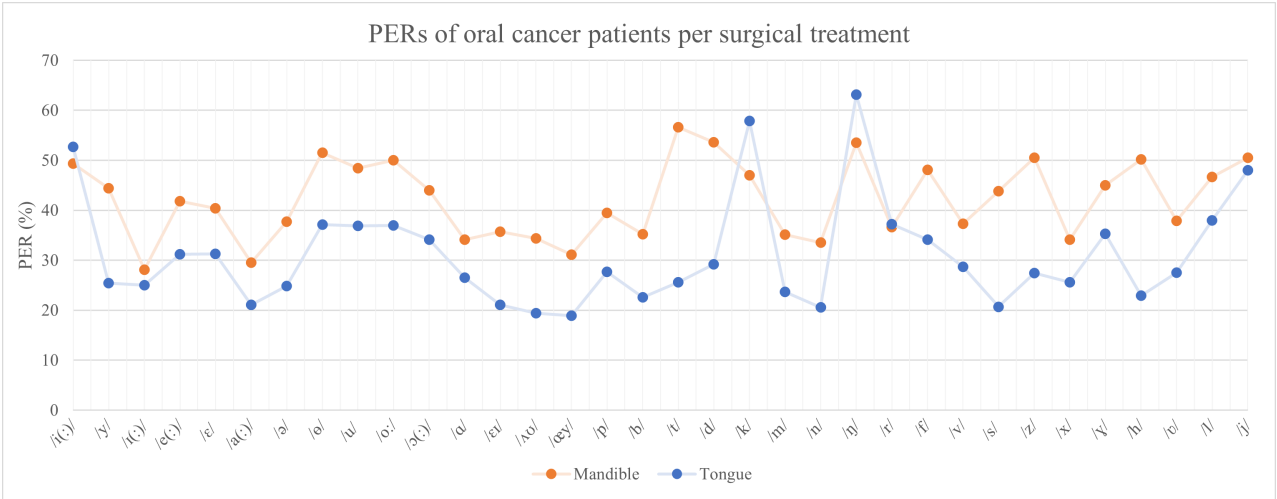


Figure 9: PER (%) results for oral cancer patients with mandible and tongue surgery.

4.3 Articulatory feature error analysis

The results of the articulatory feature error analysis are presented in Table 4. The articulatory features are grouped by both PoA and MoA, and it gives the mean AFERs and standard deviations of both speaker groups as well as overall.

Table 4: Overview of the recognition errors made on the articulatory feature level. **Blue bold** and **orange bold** numbers indicate the best and worst ASR performance for PoA and MoA per speaker group.

	AFER (%)					
	Healthy		**Patient**		**Overall**	
	Mean	*SD*	*Mean*	*SD*	*Mean*	*SD*
PoA						
Vowel	6.5	1.5	33.1	15.7	21	17.8
Bilabial	**3.2**	1.6	**30.5**	19.7	**18.1**	20
Labiodental	5.7	2	34	20.6	21.1	20.8
Alveolar	5.9	1.6	36.2	20.7	22.4	21.6
Palatal	13	3.8	**49.3**	19.4	**32.8**	23.5
Velar	5.7	2	45.2	27.1	27.3	28.2
Glottal	**15**	5.3	36.6	17.4	26.8	17
MoA						
Vowels	6.5	1.5	33.1	15.7	21	17.8
Plosives	5.7	2.2	**41**	22.8	24.9	24.5
Nasals	5.5	1.6	**29.5**	17	**18.6**	17.4
Trills	**5.3**	1.5	36.9	21.1	22.5	22.3
Fricatives	6.8	1.7	34.9	19.4	22.1	20.1
Approximants	**7**	1.4	40.8	23.5	**25.4**	24.3

4.3.1 Place of Articulation (PoA)

For the healthy speakers, the AFERs of most classes ranged between 0% and 10% (M=7.9, SD=4.4), except for the glottal /h/ and palatal /j/. This means that these two articulatory feature classes seem to have been the hardest to capture for the ASR system, followed by vowels, alveolar, labiodental and velar consonants, and finally the bilabial consonants cause the least recognition errors. When looking at the speech of the oral cancer patients, we observe that the classes have AFERs ranging between 30% and 50% (M=37.8, SD=6.8). The palatal /j/ has the highest error, followed by velars, the glottal /h/, alveolars, labiodentals, vowels, and bilabials.

4.3.2 Manner of Articulation (MoA)

For MoA, we can observe that the ASR system yielded AFERs between 5% and 10% (M=6.1, SD=0.7) for the healthy speakers, while it obtained AFERs ranging from 25% to 45% (M=36, SD=4.5)

for the speech of oral cancer patients. Even though the AFERs of all of the articulatory feature classes were overall very close (*SD*=0.7) for the healthy speakers, the analysis shows that approximants were the hardest to capture, followed by fricatives, vowels, plosives, nasals and then the trill /r/. For the oral cancer patients, the plosives were the most difficult to recognize, followed by approximants, the trill /r/, fricatives, vowels, and then nasals.

4.3.3 Effect of surgical treatment on the AFER

The results of the articulatory feature error analysis for the oral cancer patients based on their surgical treatment are shown in Figures 10 and 11. The AFER (%) is given on the y-axis and the articulatory feature classes on the x-axis.

Place of Articulation (PoA) For the patients with a mandibulectomy, the AFERs ranged from 35% to 55%, with an overall mean of 43.3% and a standard deviation of 5.7. The palatal /j/ and the glottal /h/ resulted in the highest AFERs, followed by the alveolars, velars, labiodentals, vowels, and then bilabials. For the patients with a glossectomy, the ASR system achieved AFERs ranging from 20% to 50% (*M*=32.4, *SD*=10.3). For these oral cancer patients the palatal /j/ was the most challenging as well, followed by the velars, labiodentals, vowels, alveolars, bilabials, and then the glottal /h/. Furthermore, in general better AFERs were obtained for the patients who underwent tongue surgery compared to those who underwent mandibular surgery, with the exception of the velars (see Figure 10).

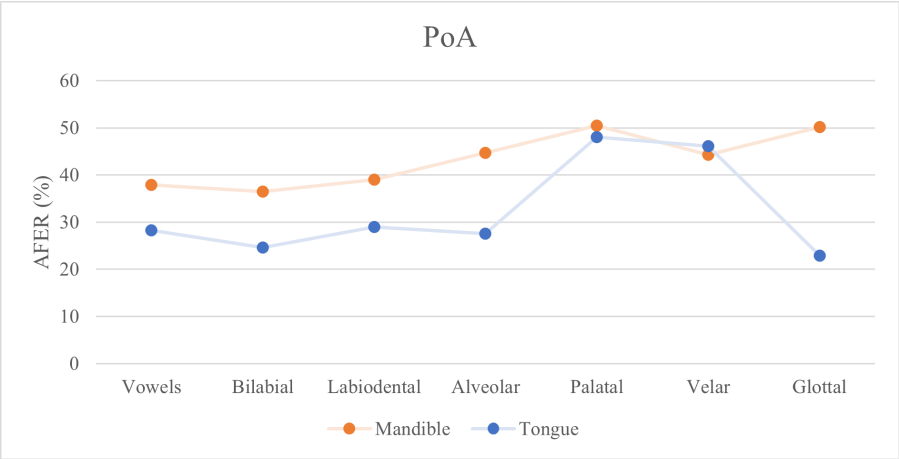


Figure 10: AFER (%) results for PoA of oral cancer patients with mandibular and tongue surgery.

Manner of Articulation (MoA) The articulatory feature error analysis for MoA yielded AFERs between 35% and 55% (*M*=41.6, *SD*=6) for the oral cancer patients with mandibular surgery, and the phoneme class that was the most challenging to recognize was that of the plosives, followed by the approximants, fricatives, vowels, the trill /r/, and finally the nasals. For the patients who have undergone a glossectomy, the ASR system achieved AFERs ranging from 20% to 40% (*M*=30.5, *SD*=5.6). In addition, it was the trill /r/ that was the most difficult to capture, followed by the approximants, plosives, vowels, fricatives and then the nasals as well. Here too, the achieved AFERs were generally lower for the patients with tongue surgery, except for the trill /r/ (see Figure 11).

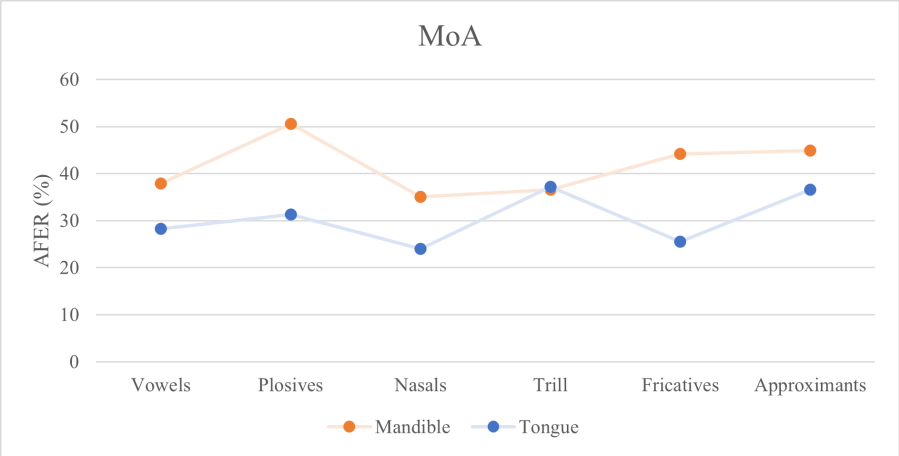


Figure 11: AFER (%) results for MoA of oral cancer patients with mandible and tongue surgery.

5 Discussion

In this chapter, the first section elaborates on the similarities between the performance of the ASR system on healthy and oral cancer speech (Section 5.1). Then the recognition errors made for the oral cancer speech are discussed in more detail (Section 5.2), followed by a discussion on the influence of the surgical treatment (Section 5.3). Finally, this chapter ends with an overview of the limitations of this study (Section 5.4).

5.1 Recognition errors in healthy and oral cancer speech

As expected, the E2E ASR system had more difficulty recognizing oral cancer speech than healthy speech, resulting in significantly higher WER scores for oral cancer speech. This confirms the findings of previous research that recognizing oral cancer speech is a challenging task for standard ASR systems. Although we can observe that our error analyses present error rates for the oral cancer speech that are certainly lower than the error rates found by Halpern et al. (2022), while our PERs for the healthy speech seem slightly higher. This is a surprising finding due to the fact that our training dataset contains a larger amount of data, including spontaneous speech.

The results of our phoneme and articulatory error analyses further reveal that the phonemes /i(ː)/, /ŋ/, and /j/ were quite challenging to recognize for the E2E ASR system in both healthy and oral cancer speech. The fact that these phonemes are recognized poorly for both types of speech indicates that it is the ASR system itself that has difficulty capturing these phonemes rather than the type of speech causing the issues. In other words, we can consider these three phonemes to be ASR-specific errors. All three of these phonemes differ in both place and manner of articulation, which means that the issues of the ASR system seem to be random and we can not find a pattern in the ASR-specific errors. A plausible explanation for the high error rates of these phonemes is that these phonemes had a lower frequency in the training data. Unfortunately, we cannot access the training data, which means we are unable to either confirm or refute this. In addition, we can observe that among these three phonemes the /ŋ/ obtains the highest PER for both healthy and oral cancer speech. This means that the /ŋ/ is the most difficult phoneme to capture for our E2E ASR system, regardless of the type of speech.

Furthermore, when we look at the articulatory feature error analysis for PoA, the results show that bilabials are most often correctly recognized for healthy speech as well as for oral cancer speech. For the oral cancer patients, this could be explained by the fact that the tongue is not involved in the articulation of these phonemes, resulting in better ASR performance for these phonemes.

5.2 Recognition errors specific to oral cancer speech

In the previous section we have given an overview of the similarities of the ASR performance on healthy and oral cancer speech, though in this section we wish to take a closer look at the differences, i.e. the recognition errors made for oral cancer speech specifically. The results of the phoneme error analysis revealed that the phonemes /i(ː)/, /ŋ/, /k/, and /j/ had a PER exceeding the 45% threshold and they are therefore considered to be poorly recognized by the ASR system. However, in Section 5.1 we have argued that the high error rates for /i(ː)/, /ŋ/, and /j/ were due to the ASR system generally not capturing these phonemes well. This indicates that only the phoneme /k/ is relatively more challenging to recognize for oral cancer speech than it is for healthy speech. This is in line with the findings of Borggreven et al. (2005), who reported that the /k/ was frequently misrecognized in Dutch oral cancer

speech. In addition, de Bruijn et al. (2009) state that /k/ is one of the phonemes that functions as a predictor of Dutch oral cancer speech.

Furthermore, this outcome is supported by the articulatory feature error analysis for the PoA, as it shows that velars are the second most challenging for the ASR system to capture in oral cancer speech, which was also reported by Halpern et al. (2022) for their E2E ASR system. Moreover, the ASR performance based on PoA is quite similar for the two speaker groups, except for the fact that velars are recognized quite well for healthy speakers and not for oral cancer patients. Therefore, the only articulatory feature class that is recognized comparatively worse for oral cancer speech than for healthy speech is that of the velars. These results can be explained by the fact that the oral cancer patients all have reduced tongue motility to a certain degree, while these phonemes require an adequate motility of the tongue and the velum (de Bruijn et al., 2009). In addition, since the /ŋ/ is poorly recognized by the ASR system in general, and /x/ and /ɣ/ are captured relatively well, this means that the /k/ is mainly responsible for this finding.

Even though the /j/ and /ŋ/ were poorly recognized in healthy speech as well, the PoA analysis did reveal that both palatals and velars were the most challenging to capture in oral cancer speech. This also corresponds to the results reported by Halpern et al. (2022), who found that palatals and velars obtained the highest AFERs for their E2E ASR system. The high PER of the vowel /i(ː)/, however, seems to contradict our PoA analysis for oral cancer speech, since vowels achieved the second best ASR performance. It must be kept in mind though, that the articulatory feature class of vowels comprises a rather large number of phonemes, which causes the vowels as an entire class not to be impacted a lot by the PER of a single phoneme. Moreover, it seems very probable that this mismatch between the PER and AFER is due to the /i(ː)/ being poorly recognized in healthy speech as well, meaning it is an ASR-specific error.

The articulatory feature analysis for the MoA further supports the high PERs of the phonemes /j/ and /k/, since both plosives and approximants were misrecognized most frequently in oral cancer speech. This is in accordance with the literature, as previous research has shown that plosives are important predictors for oral cancer speech because they are impacted the most by the surgical treatment (e.g. Bressmann et al., 2004, 2009; Borggreven et al., 2005; de Bruijn et al., 2009). Similarly, Halpern et al. (2022) reported that plosives and approximants are among the articulatory feature classes to be captured with more difficulty by standard E2E ASR systems that are pre-trained on healthy speech. However, the high PER scores of the /i(ː)/ and /ŋ/ are seemingly in contrast to the findings of the articulatory feature analysis for MoA, as both the vowels and the nasals yielded the best AFER results. Nevertheless, even though the vowels in our research were relatively well captured for both MoA and PoA, previous research has also found the /i(ː)/ to be affected in oral cancer speech (Whitehill et al., 2006; de Bruijn et al., 2009). Regarding the /ŋ/, on the other hand, it simply seems to be the case that the PoA was more dominant than the MoA.

Even though previous research has indicated that, besides plosives, sibilants are impacted in oral cancer speech as well (e.g. Borggreven et al., 2005; Laaksonen et al., 2011), we did not observe that the ASR system had more difficulty with recognizing sibilants. In fact, the production of /s/ and /z/ was captured relatively well. It must be mentioned, however, that we were unable to report on the postalveolar sibilants /ʃ/ and /ʒ/ due to the small amount of occurrences, which could have influenced our results. Nevertheless, Halpern et al. (2022) presented similar outcomes regarding sibilants and they suggested that it might be due to the fact that sibilants can be considered noise, meaning that

loss of information would make less of a difference for ASR systems than it would for other types of sounds. In addition, previous research used human listeners to assess the oral cancer speech, whereas this study used an ASR system, which relies on a language model.

Even though the /i(ː)/ is the only vowel with a PER over 45% and vowels are captured relatively well for oral cancer speech, we can observe that three other vowels had PERs higher than 40%, namely the /ɵ/, /oː/, and /u/. This seems to indicate that, with the exception of /i(ː)/, the vowels that are affected most in oral cancer speech are central and back vowels that are rounded and realised with a relatively high degree of constriction by the tongue. Although research on Dutch oral cancer speech does mention a compressed vowel space area (de Bruijn et al., 2009), it does not mention these vowels specifically. For English oral cancer speech however, Halpern et al. (2022) also found the production of /u/ to cause difficulty for the E2E ASR system.

In summary, the answer to our first research question is that especially the production of the phoneme /k/ is difficult for standard E2E ASR systems to capture in Dutch oral cancer speech compared to healthy speech. The articulatory feature error analyses for PoA and MoA confirm this finding, as both plosives and velars are among the articulatory feature classes that are misrecognized most frequently.

5.3 Influence of the type of surgical treatment

The statistical analysis revealed that there was no significant difference between the WER results of patients who underwent a (partial) glossectomy and the speech of the patients who underwent a mandibulectomy. It must be mentioned, however, that this insignificance is very likely the result of the small sample size. Nevertheless, in general the speech of oral cancer patients who had mandibular surgery did yield higher WERs than the speech of patients with a glossectomy, and this can be explained by the fact that a mandibulectomy impacts more articulators (mandible and tongue) than a glossectomy (tongue only) (Matsui et al., 2007). In line with this, we can observe that the two patients with the highest WER scores had received surgery to their mandible. However, the patient with the best WER had undergone a mandibulectomy as well, which seems to contradict our statement on the greater impact of a mandibulectomy in comparison to a glossectomy. Nevertheless, it is possible that this patient's tumour was less severe and that she therefore received surgical treatment that was less invasive than the treatment of the other patients. This in turn could have caused the speech of this patient to be less impaired and would thus lead to a better ASR performance. In addition, this high WER instance could be explained by other factors, as previous research has suggested that the resection site (Logemann et al., 1993; Borggreven et al., 2005) or reconstruction technique (Konstantinović and Dimić, 1998) can influence speech intelligibility and therefore the ASR performance.

In accordance with the WER results, the phoneme error analysis revealed that better ASR performance was achieved for patients with a glossectomy than for patients with a mandibulectomy with the exception of the phonemes /i(ː)/, /k/, and /ŋ/. These were also the only three phonemes that had a PER over 50% for the patients with tongue surgery. As mentioned before, we can explain the PERs of the /i(ː)/ and /ŋ/, since these phonemes were generally poorly captured by the ASR system. Regarding the /k/, it is interesting to notice that it is a velar and that it requires the airstream to be fully obstructed, in contrast to the other velars /x/ and /ɣ/. It seems to be the case that raising the tongue to the lowered velum is the most difficult articulatory movement to make for patients with a glossectomy. Although the /i(ː)/ was poorly recognized by the ASR system in general, it also requires the tongue to be raised. However, it is actually a front vowel rather than a back vowel, which is in contrast to the PoA of the

/k/. Nevertheless, as mentioned in Section 5.2 our findings are in agreement with previous research (de Bruijn et al., 2009).

For the patients with mandibular surgery, the phoneme error analysis revealed that the phonemes eliciting PERs over 50% are /ɵ/, /oː/, /t/, /d/, /ŋ/, /z/, /h/, and /j/. Of these phonemes the /t/ and /d/ were misrecognized most frequently, which is interesting as these phonemes have the same PoA and MoA and thus differ solely in voicing. Moreover, this is in line with the observations of Borggreven et al. (2005) that the alveolar plosives are difficult to produce for Dutch oral cancer patients.

The findings discussed above are confirmed by the articulatory feature error analysis. The results show that for both PoA and MoA the ASR system again had better performance for the speech of patients with tongue surgery compared to the speech of the patients with mandibular surgery. The speech of the patients with tongue surgery did yield higher AFERs for velars, which is in line with the PER results. Moreover, for both patient groups the palatal /j/ achieved the worst ASR performance for PoA. It could, however, be the result of the low number of occurrences and/or the fact that the whole articulatory feature class is represented by a single phoneme. For MoA, only the trill /r/ elicited a higher AFER score for the patients with tongue surgery than for patients who underwent a mandibulectomy, although the AFERs of the two groups only differ slightly. In addition, the trill is the articulatory feature class that scored the highest AFER for the patients with tongue surgery. This is not surprising, since the production of this phoneme is completely dependent on the (tip of the) tongue, which is (partly) removed after a glossectomy. It must be mentioned, however, that the /r/ did not obtain a particularly high PER compared to other phonemes and that the high AFER score is probably caused by the fact that the entire class is represented by a single phoneme. For the patients with mandibular surgery, on the other hand, the plosives were the most difficult to capture, which corresponds to previous literature (e.g. Bressmann et al., 2004, 2009; Halpern et al., 2022). In addition, approximants were the second hardest articulatory feature class to be captured for both patient groups.

The articulatory feature error analysis for MoA further revealed that the nasals yielded the best AFER scores in both patient groups, which contradicts the fact that the /ŋ/ is among the phonemes to yield the highest PER score for both groups. As mentioned in Section 5.2, a plausible explanation for this outcome is that for this particular phoneme the PoA is more dominant than the MoA, and thus results in the velar /ŋ/ being the only nasal to be captured relatively poorly by the ASR system.

Furthermore, the observations made in Section 4.2.1 regarding patients 04 and 05 seem to confirm that velars are particularly challenging to capture for the ASR system when the oral cancer patient underwent a glossectomy. This is due to the fact that patient 04, who underwent a glossectomy, had the highest PER for all of the velars, while the speech of patient 05, who had mandibular surgery, obtained the lowest PERs for the velar phonemes. Additionally, the finding that patient 02 had the lowest PER scores for almost all vowels follows logically from the fact that he had a mandibulectomy. To elaborate, the jaw plays an important role in the production of vowels, especially in influencing the height of vowels (Mooshammer et al., 2007), which indicates that it is more challenging to produce certain vowels for oral cancer patients with mandibular surgery in comparison to oral cancer patients with tongue surgery.

To summarize, in answer to our second research question we can argue that there seems to be a slight influence of the type of surgical treatment on the ASR performance. The speech of patients who have undergone a mandibulectomy yielded higher recognition error rates than the speech of patients with

tongue surgery, although our statistical analysis revealed that the difference between the two patient groups failed to reach significance. In addition, while the speech of the glossectomy patients mostly obtained recognition errors for velars and the trill /r/, the speech of the patients with a mandibulectomy yielded the highest recognition error rates for the palatal /j/ and the glottal /h/ for PoA and plosives for MoA.

5.4 Limitations and future recommendations

It is important to acknowledge that this study has several limitations. First of all, even though ASR of oral cancer speech is considered to be a low-resource recognition task, the amount of data used in this study was still very limited. Several phonemes had to be excluded from the analysis, as the number of occurrences was too low and we were unable to provide a full analysis of all the phonemes in the Dutch language. Therefore, it is recommended for future research on Dutch oral cancer speech to use stimuli that consist of enough occurrences of every phoneme that exists in Dutch to be able to perform reliable (error) analyses. Furthermore, the results regarding the influence of the type of surgery should be interpreted with caution, since both groups contained speech data of only three patients. The small amount of data prevents us from observing the outliers, which makes it difficult to draw valid conclusions and challenging to avoid overgeneralization. Thus, we encourage researchers to gather more data for both patient groups in order to gain a deeper understanding of what types of errors oral cancer patients make after receiving different types of surgical treatments.

Secondly, it must be mentioned that the oral cancer speech dataset does not accurately represent Dutch oral cancer speakers, due to the fact that all of the participants come from the northern regions of the Netherlands. Naturally, the way people speak in the North does not correspond to the way people speak in the South, and we therefore suggest to include speakers from all of the regions in the Netherlands in future experiments.

Thirdly, our dataset only contains read speech rather than spontaneous speech. When people are reading out loud they pay more attention to how they speak, causing read speech to be significantly different from spontaneous speech in both acoustic and linguistic terms (Nakamura et al., 2008). Since the very purpose of ASR systems developed for oral cancer speech is to ease communication in the daily life of the patients, read speech does not give an accurate representation of how oral cancer patients would interact with ASR systems. In addition, this study did not take into account the impact of phonological processes such as assimilation. Processes like assimilation occur very frequently in spontaneous speech, and it is therefore very important to keep these processes in mind. Therefore, we would recommend future researchers to either use spontaneous speech only or to augment the dataset with spontaneous speech.

Lastly, we believe it would be interesting for future research to investigate the influence of the training data on the ASR performance for oral cancer speech. Thus, we suggest to conduct experiments in which the training data is augmented with oral cancer speech or consists solely of oral cancer speech by adapting retraining methods similar to the ones employed by Halpern et al. (2022).

6 Conclusion

Over the last couple of years, the use of automatic speech recognition in daily life has gained popularity, as it makes people's lives more convenient. However, current ASR systems are typically trained on standard healthy speech, resulting in poor ASR performance for people with impaired speech, such as oral cancer patients. Since ASR systems could greatly improve the quality of life of oral cancer patients, it is therefore necessary to develop ASR systems specifically for oral cancer speech. In order to further the development of such ASR systems, we investigated the type of recognition errors made for Dutch oral cancer speech.

This study set out to gain insights into the recognition errors of a standard E2E ASR system that was pre-trained on healthy speech when it recognized oral cancer speech in Dutch. In addition, we investigated whether the type of surgical treatment that the oral cancer patients received would influence the ASR performance. In order to answer our research questions, we used a Dutch speech database containing both healthy and oral cancer speech and ran it through an ESPnet-based model that was pre-trained on healthy Dutch speech. Then we performed an extensive error analysis on the word, phoneme, and articulatory feature level.

In answer to our first research question, we found that particularly the production of the phoneme /k/ was challenging to capture for the standard ASR system. This is in line with previous research stating that the /k/ is an important predictor of Dutch oral cancer speech (de Bruijn et al., 2009). In addition, our error analyses for place and manner of articulation were in agreement with this finding as well, as we found that plosives and velars elicited the highest and second highest recognition error rates in oral cancer speech. Similar results regarding the plosives were reported in several studies (e.g. Borggreven et al., 2005; Bressmann et al., 2004; de Bruijn et al., 2009), though velars have only been specifically mentioned by Halpern et al. (2022). Furthermore, we did not find the sibilants to be comparatively more challenging to capture in oral cancer speech than in healthy speech, which seemingly contradicts existing literature on oral cancer speech (e.g. Laaksonen et al., 2011; Borggreven et al., 2005), although Halpern et al. (2022) did not find results like this for English oral cancer either. It must be kept in mind, however, that the sibilants /ʃ/ and /ʒ/ were excluded from our analysis, which could have influenced our results. With regard to the vowels, we expected to find /a/ and /u/ to be relatively poorly recognized for oral cancer speech based on the findings of Halpern et al. (2022). However, the /a/ was captured comparatively well and although the /u/ was among the vowels to yield the highest PER, it did not exceed the threshold we set for oral cancer speech. We did find the /i(ː)/ to be particularly poorly recognized, although this was the case in healthy speech as well and we therefore consider it to be an ASR-specific error.

Regarding our second research question, we found that the speech of patients who underwent a mandibulectomy elicited higher recognition error rates than the speech of patients who had a (partial) glossectomy. This is in accordance with our hypothesis, and can be explained by the fact that mandibular surgery impacts more articulators than tongue surgery (Matsui et al., 2007). More specifically, the phoneme error analysis revealed that the ASR system had the most difficulty with the production of the phonemes /t/ and /d/ in the speech of patients with mandibular surgery, which is in agreement with the findings of Borggreven et al. (2005), who found that alveolar plosives are challenging to produce for Dutch oral cancer patients. In addition, our articulatory feature error analysis revealed that plosives yielded the highest error rates for speech of patients with a mandibulectomy. For the patients with tongue surgery, it was the production of the phoneme /k/ that caused the most

difficulty for the ASR system, which is confirmed by the fact that velars obtained the highest error rates for patients with a glossectomy.

In conclusion, the outcomes of our study are generally in accordance with the existing literature on the characteristics of oral cancer speech (e.g. Borggreven et al., 2005; Halpern et al., 2022). However, the amount of data in our study was very limited and caution should be taken regarding the generality of our results. Therefore, future research on the recognition errors of standard ASR systems for Dutch oral cancer speech is crucial for the development of ASR systems that can accurately recognize the speech of Dutch oral cancer patients. Our research was the first step.

Bibliography

Alzubaidi, L., Zhang, J., Humaidi, A. J., Al-Dujaili, A., Duan, Y., Al-Shamma, O., Santamaría, J., Fadhel, M. A., Al-Amidie, M., and Farhan, L. (2021). Review of deep learning: concepts, CNN architectures, challenges, applications, future directions. *Journal of Big Data*, 8(1).

Bahdanau, D., Chorowski, J., Serdyuk, D., Brakel, P., and Bengio, Y. (2016). End-to-end attention-based large vocabulary speech recognition. In *2016 IEEE international conference on acoustics, speech and signal processing (ICASSP)*, pages 4945–4949. IEEE.

Bender, B. K., Cannito, M. P., Murry, T., and Woodson, G. E. (2004). Speech Intelligibility in Severe Adductor Spasmodic Dysphonia. *Journal of Speech, Language, and Hearing Research*, 47(1):21–32.

Berger, B., Waterman, M. S., and Yu, Y. W. (2021). Levenshtein Distance, Sequence Comparison and Biological Database Search. *IEEE Transactions on Information Theory*, 67(6):3287–3294.

Bernard, M. and Titeux, H. (2021). Phonemizer: Text to phones transcription for multiple languages in python. *Journal of Open Source Software*, 6(68):3958.

Borggreven, P. A., Verdonck-de Leeuw, I., Langendijk, J. A., Doornaert, P., Koster, M. N., de Bree, R., and Leemans, C. R. (2005). Speech outcome after surgical treatment for oral and oropharyngeal cancer: A longitudinal assessment of patients reconstructed by a microvascular flap. *Head Neck*, 27(9):785–793.

Bressmann, T., Jacobs, H., Quintero, J., and Irish, J. C. (2009). Speech outcomes for partial glossectomy surgery: Measures of speech articulation and listener perception indicateurs de la parole pour une glossectomie partielle: Mesures de l'articulation de la parole et de la perception des auditeurs. *Head and Neck Cancer*, 33(4):204–210.

Bressmann, T., Sader, R., Whitehill, T. L., and Samman, N. (2004). Consonant intelligibility and tongue motility in patients with partial glossectomy. *Journal of Oral and Maxillofacial Surgery*, 62(3):298–303.

Calvo, I., Tropea, P., Viganò, M., Scialla, M., Cavalcante, A. B., Grajzer, M., Gilardone, M., and Corbo, M. (2020). Evaluation of an Automatic Speech Recognition Platform for Dysarthric Speech. *Folia Phoniatrica et Logopaedica*, 73(5):432–441.

Chan, W., Jaitly, N., Le, Q., and Vinyals, O. (2016). Listen, attend and spell: A neural network for large vocabulary conversational speech recognition. In *2016 IEEE international conference on acoustics, speech and signal processing (ICASSP)*, pages 4960–4964. IEEE.

Cho, K., Van Merriënboer, B., Bahdanau, D., and Bengio, Y. (2014). On the properties of neural machine translation: Encoder-decoder approaches. *arXiv preprint arXiv:1409.1259*.

Christensen, H., Aniol, M., Bell, P., Green, P. D., Hain, T., King, S., and Swietojanski, P. (2013). Combining in-domain and out-of-domain speech data for automatic recognition of disordered speech. In *INTERSPEECH*, pages 3642–3645.

Chung, J., Gulcehre, C., Cho, K., and Bengio, Y. (2014). Empirical evaluation of gated recurrent neural networks on sequence modeling. *arXiv preprint arXiv:1412.3555*.

de Bruijn, M. J., ten Bosch, L., Kuik, D. J., Quené, H., Langendijk, J. A., Leemans, C. R., and Verdonck-de Leeuw, I. M. (2009). Objective Acoustic-Phonetic Speech Analysis in Patients Treated for Oral or Oropharyngeal Cancer. *Folia Phoniatrica et Logopaedica*, 61(3):180–187.

Deng, K., Cheng, G., Yang, R., and Yan, Y. (2022). Alleviating ASR Long-Tailed Problem by Decoupling the Learning of Representation and Classification. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 30:340–354.

Enderby, P. (2013). Disorders of communication: dysarthria. In Barnes, M. P. and Good, D. C., editors, *Neurological Rehabilitation*, volume 110, chapter 22, pages 273–282. Elsevier.

Epstein, J. B., Emerton, S., Kolbinson, D. A., Le, N. D., Phillips, N., Stevenson-Moore, P., and Osoba, D. (1999). Quality of life and oral function following radiotherapy for head and neck cancer. *Head Neck*, 21(1):1–11.

Fant, G. (1981). The source filter concept in voice production. *STL-QPSR*, 1(1981):21–37.

Graves, A. and Jaitly, N. (2014). Towards end-to-end speech recognition with recurrent neural networks. In *International conference on machine learning*, pages 1764–1772. PMLR.

Graves, A., Mohamed, A.-r., and Hinton, G. (2013). Speech recognition with deep recurrent neural networks. In *2013 IEEE international conference on acoustics, speech and signal processing*, pages 6645–6649. IEEE.

Gulati, A., Qin, J., Chiu, C.-C., Parmar, N., Zhang, Y., Yu, J., Han, W., Wang, S., Zhang, Z., Wu, Y., et al. (2020). Conformer: Convolution-augmented transformer for speech recognition. *arXiv preprint arXiv:2005.08100*.

Halpern, B. M., Feng, S., van Son, R., van den Brekel, M., and Scharenborg, O. (2022). Low-resource automatic speech recognition and error analyses of oral cancer speech. *Speech Communication*.

Halpern, B. M., van Son, R., van den Brekel, M., and Scharenborg, O. (2020). Detecting and analysing spontaneous oral cancer speech in the wild. In *INTERSPEECH*, volume 21, pages 4826–4830.

Harvill, J., Issa, D., Hasegawa-Johnson, M., and Yoo, C. (2021). Synthesis of new words for improved dysarthric speech recognition on an expanded vocabulary. In *ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 6428–6432. IEEE.

Hermann, E. and Doss, M. M. (2020). Dysarthric speech recognition with lattice-free MMI. In *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 6109–6113. IEEE.

Hochreiter, S. and Schmidhuber, J. (1997). Long short-term memory. *Neural computation*, 9(8):1735–1780.

Jurafsky, D. and Martin, J. (2020). *Speech and Language Processing*. Els autors, 3 edition.

Kappert, K. D. R., van Alphen, M. J. A., Smeele, L. E., Balm, A. J. M., and van der Heijden, F. (2019). Quantification of tongue mobility impairment using optical tracking in patients after receiving primary surgery or chemoradiation. *PLOS ONE*, 14(8):e0221593.

Koenecke, A., Nam, A., Lake, E., Nudell, J., Quartey, M., Mengesha, Z., Toups, C., Rickford, J. R., Jurafsky, D., and Goel, S. (2020). Racial disparities in automated speech recognition. *Proceedings of the National Academy of Sciences*, 117(14):7684–7689.

Konstantinović, V. and Dimić, N. (1998). Articulatory function and tongue mobility after surgery followed by radiotherapy for tongue and floor of the mouth cancer patients. *British Journal of Plastic Surgery*, 51(8):589–593.

Kreeft, A. M., van der Molen, L., Hilgers, F. J., and Balm, A. J. (2009). Speech and swallowing after surgical treatment of advanced oral and oropharyngeal carcinoma: a systematic review of the literature. *European Archives of Oto-Rhino-Laryngology*, 266(11):1687–1698.

Laaksonen, J.-P., Rieger, J., Harris, J., and Seikaly, H. (2011). A longitudinal acoustic study of the effects of the radial forearm free flap reconstruction on sibilants produced by tongue cancer patients. *Clinical Linguistics Phonetics*, 25(4):253–264.

Lazarus, C. L., Husaini, H., Anand, S. M., Jacobson, A. S., Mojica, J. K., Buchbinder, D., and Urken, M. L. (2013). Tongue Strength as a Predictor of Functional Outcomes and Quality of Life after Tongue Cancer Surgery. *Annals of Otology, Rhinology Laryngology*, 122(6):386–397.

Lazarus, C. L., Wall, L. R., Ward, E. C., and Yiu, E. (2014). Speech and swallowing following oral, oropharyngeal, and nasopharyngeal cancers. In Ward, E. C. and As-Brooks, C. J. v., editors, *Head and neck cancer: treatment, rehabilitation, and outcomes*. Plural Publishing, second edition.

Lindsay, H., Tröger, J., and König, A. (2021). Language Impairment in Alzheimer's Disease—Robust and Explainable Evidence for AD-Related Deterioration of Spontaneous Speech Through Multilingual Machine Learning. *Frontiers in Aging Neuroscience*, 13.

Logemann, J. A., Pauloski, B. R., Rademaker, A. W., McConnel, F. M. S., Heiser, M. A., Cardinale, S., Shedd, D., Stein, D., Beery, Q., Johnson, J., and Baker, T. (1993). Speech and Swallow Function After Tonsil/Base of Tongue Resection With Primary Closure. *Journal of Speech, Language, and Hearing Research*, 36(5):918–926.

Maier, A., Haderlein, T., Stelzle, F., Nöth, E., Nkenke, E., Rosanowski, F., Schützenberger, A., and Schuster, M. (2010). Automatic Speech Recognition Systems for the Evaluation of Voice and Speech Disorders in Head and Neck Cancer. *EURASIP Journal on Audio, Speech, and Music Processing*, 2010:1–7.

Matsui, Y., Ohno, K., Yamashita, Y., and Takahashi, K. (2007). Factors influencing postoperative speech function of tongue cancer patients following reconstruction with fasciocutaneous/myocutaneous flaps—a multicenter study. *International Journal of Oral and Maxillofacial Surgery*, 36(7):601–609.

McFee, B., Raffel, C., Liang, D., Ellis, D. P. W., McVicar, M., Battenberg, E., and Nieto, O. (2015). librosa: Audio and music signal analysis in python. In *Proceedings of the 14th python in science conference*, pages 18–25.

Miao, H., Cheng, G., Zhang, P., Li, T., and Yan, Y. (2019). Online hybrid ctc/attention architecture for end-to-end speech recognition. In *Interspeech*, pages 2623–2627.

Mooshammer, C., Hoole, P., and Geumann, A. (2007). Jaw and Order. *Language and Speech*, 50(2):145–176.

Muhammad, G., Mesallam, T. A., Malki, K. H., Farahat, M., Alsulaiman, M., and Bukhari, M. (2011). Formant analysis in dysphonic patients and automatic Arabic digit speech recognition. *BioMedical Engineering OnLine*, 10(1):41.

Nakamura, M., Iwano, K., and Furui, S. (2008). Differences between acoustic characteristics of spontaneous and read speech and their effects on speech recognition performance. *Computer Speech Language*, 22(2):171–184.

Nederlandse Taalunie (1998-2004). Corpus Gesproken Nederlands (CGN) (Version 2.0.3). [Dataset].

Paszke, A., Gross, S., Chintala, S., Chanan, G., Yang, E., DeVito, Z., Lin, Z., Desmaison, A., Antiga, L., and Lerer, A. (2017). Automatic differentiation in pytorch. In *Proceedings of The future of gradientbased machine learning software and techniques (Autodiff) in the twenty-ninth annual conference on neural information processing systems (NIPS)*.

Povey, D., Ghoshal, A., Boulianne, G., Burget, L., Glembek, O., Goel, N., Hannemann, M., Motlicek, P., Qian, Y., Schwarz, P., et al. (2011). The kaldi speech recognition toolkit. In *IEEE 2011 workshop on automatic speech recognition and understanding*, number CONF. IEEE Signal Processing Society.

Qin, Y., Lee, T., Kong, A. P. H., and Law, S. (2018). Application of automatic speech recognition (ASR) techniques for automatic speech assessment in people with aphasia. *Frontiers in Human Neuroscience*, 12.

Qin, Y., Wu, Y., Lee, T., and Kong, A. P. H. (2020). An end-to-end approach to automatic speech assessment for cantonese-speaking people with aphasia. *Journal of Signal Processing Systems*, 92(8):819–830.

Ravanelli, M., Brakel, P., Omologo, M., and Bengio, Y. (2018). Light Gated Recurrent Units for Speech Recognition. *IEEE Transactions on Emerging Topics in Computational Intelligence*, 2(2):92–102.

Rentschler, G. and Mann, M. (1980). The effects of glossectomy on intelligibility of speech and oral perceptual discrimination. *Journal of Oral Surgery (American Dental Association: 1965)*, 38(5):348–354.

Rullo, R., Di Maggio, D., Festa, V., and Mazzarella, N. (2009). Speech assessment in cleft palate patients: A descriptive study. *International Journal of Pediatric Otorhinolaryngology*, 73(5):641–644.

Safaiean, A., Jalilevand, N., Ebrahimipour, M., Asleshirin, E., and Hiradfar, M. (2017). Speech intelligibility after repair of cleft lip and palate. *Medical Journal of the Islamic Republic of Iran*, 31(1):500–504.

Saravanan, V., Ranganathan, V., Gandhi, A., and Jaya, V. (2016). Speech outcome in oral cancer patients - pre- and post-operative evaluation: A cross-sectional study. *Indian Journal of Palliative Care*, 22(4):499–503.

Seikel, A., Drumright, D., and Hudock, D. (2019). *Anatomy Physiology for Speech, Language, and Hearing*. Plural Publishing, Inc., 6 edition.

Shahamiri, S. R. (2021). Speech Vision: An End-to-End Deep Learning-Based Dysarthric Automatic Speech Recognition System. *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, 29:852–861.

Sharma, H. V. and Hasegawa-Johnson, M. (2013). Acoustic model adaptation using in-domain background models for dysarthric speech recognition. *Computer Speech & Language*, 27(6):1147–1162.

Shield, K. D., Ferlay, J., Jemal, A., Sankaranarayanan, R., Chaturvedi, A. K., Bray, F., and Soerjomataram, I. (2016). The global incidence of lip, oral cavity, and pharyngeal cancers by subsite in 2012. *CA: A Cancer Journal for Clinicians*, 67(1):51–64.

Shin, Y. S., Koh, Y. W., Kim, S.-H., Jeong, J. H., Ahn, S., Hong, H. J., and Choi, E. C. (2012). Radiotherapy Deteriorates Postoperative Functional Outcome After Partial Glossectomy With Free Flap Reconstruction. *Journal of Oral and Maxillofacial Surgery*, 70(1):216–220.

Sugomori, Y., Kaluza, B., Soares, F. M., and Souza, A. M. F. (2017). *Deep Learning*. Packt Publishing.

Takatsu, J., Hanai, N., Suzuki, H., Yoshida, M., Tanaka, Y., Tanaka, S., Hasegawa, Y., and Yamamoto, M. (2017). Phonologic and Acoustic Analysis of Speech Following Glossectomy and the Effect of Rehabilitation on Speech Outcomes. *Journal of Oral and Maxillofacial Surgery*, 75(7):1530–1541.

Tatman, R. and Kasten, C. (2017). Effects of Talker Dialect, Gender & Race on Accuracy of Bing Speech and YouTube Automatic Captions. In *Interspeech*, pages 934–938.

Tokui, S., Oono, K., Hido, S., and Clayton, J. (2015). Chainer: a next-generation open source framework for deep learning. In *Proceedings of workshop on machine learning systems (LearningSys) in the twenty-ninth annual conference on neural information processing systems (NIPS)*, volume 5, pages 1–6.

van der Molen, L., van Rossum, M. A., Burkhead, L. M., Smeele, L. E., and Hilgers, F. J. M. (2008). Functional outcomes and rehabilitation strategies in patients treated with chemoradiotherapy for advanced head and neck cancer: a systematic review. *European Archives of Oto-Rhino-Laryngology*, 266(6):889–900.

van der Molen, L., van Rossum, M. A., Jacobi, I., van Son, R. J., Smeele, L. E., Rasch, C. R., and Hilgers, F. J. (2012). Pre- and Posttreatment Voice and Speech Outcomes in Patients With Advanced Head and Neck Cancer Treated With Chemoradiotherapy: Expert Listeners' and Patient's Perception. *Journal of Voice*, 26(5):664.e25–664.e33.

Watanabe, S., Hori, T., Karita, S., Hayashi, T., Nishitoba, J., Unno, Y., Soplin, N. E. Y., Heymann, J., Wiesner, M., Chen, N., Renduchintala, A., and Ochiai, T. (2018). ESPnet: End-to-End Speech Processing Toolkit. In *INTERSPEECH*, pages 2207–2211.

Watanabe, S., Hori, T., Kim, S., Hershey, J. R., and Hayashi, T. (2017). Hybrid ctc/attention architecture for end-to-end speech recognition. *IEEE Journal of Selected Topics in Signal Processing*, 11(8):1240–1253.

Whitehill, T. L., Ciocca, V., Chan, J. C., and Samman, N. (2006). Acoustic analysis of vowels following glossectomy. *Clinical Linguistics Phonetics*, 20(2-3):135–140.

Windrich, M., Maier, A., Kohler, R., Nöth, E., Nkenke, E., Eysholdt, U., and Schuster, M. (2008). Automatic Quantification of Speech Intelligibility of Adults with Oral Squamous Cell Carcinoma. *Folia Phoniatrica et Logopaedica*, 60(3):151–156.

Yilmaz, E., Ganzeboom, M., Cucchiarini, C., and Strik, H. (2017). Multi-stage DNN training for automatic recognition of dysarthric speech.

Zhang, Q., Lu, H., Sak, H., Tripathi, A., McDermott, E., Koo, S., and Kumar, S. (2020). Transformer transducer: A streamable speech recognition model with transformer encoders and rnn-t loss. In *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 7829–7833. IEEE.

Zhang, Y. (2022). *Mitigating bias against Non-native accents*. PhD thesis, Delft University of Technology.

A Data agreement

Data agreement regarding data sharing with students, research assistants and interns

The following agreement concerns data collected within project '*Articulation and coordination of speech after treatment for oral cancer*', which is a tripartite collaboration between the University Medical Center Groningen (prof. dr. M.J.H. Witjes), the Netherlands Cancer Institute (prof. dr. M.W.M. den Brekel & dr. R.J.J.H. van Son), and the University of Groningen (prof. dr. Martijn Wieling).

The agreement is between the student, research assistant or intern (hereinafter 'student'):
……

and the project representative (hereinafter 'representative'):
……

The goal of the agreement is to protect project data that will be shared with the student for the following purpose:
……
……
……

By signing the agreement, the student acknowledges that they are aware of the following:

- The data (hereinafter 'data') includes the acoustic recordings of oral cancer patients and healthy control speakers, and the minimal demographic and medical information (hereinafter 'metadata') required for the purpose of the research. All data shared is confidential and should be treated as such. Data will remain anonymous at all times.
- Metadata will be shared only if that is necessary for the purposes of the student's work with the data. If demographic and medical information is needed, only the necessary subparts will be shared.
- Data can be used by the student for the above-described purposes until the end of the student's project. The representative is responsible for revoking access to the data after the end of the student's project.
- Data can be accessed only through the university workplace, either through the UWP (Windows) or LWP (Linux) servers. The representative makes sure that the data can be accessed by the students.
- Data should remain on the university servers and should not be downloaded or transferred to a different machine. That includes, but is not limited to, uploading the data on cloud storage, external hard drive, or personal computers.
- Data should not be shown or demonstrated during presentations (including course purposes), unless explicit permission is obtained from the representative beforehand.
- If any of the above terms are violated, use of data is no longer allowed, and the representative will revoke access to the data. The representative can revoke access to the data at any time without any justification.

The agreement was signed on …… in ……., in two exemplars. The representative and the student each received a signed copy.

Student's signature: ……

Representative's signature: ……

B Sentences read by participants

LIT1: Papa en Marloes

1. Papa en Marloes staan op het station.
2. Ze wachten op de trein.
3. Ze wachten op de trein.
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.

LIT2: 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.

LIT3: 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.

LIT4: Els gaat naar de 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.

LIT5: 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.

LIT6: 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.

NEWS1

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.

NEWS2

1. Door haar bekendheid kreeg Moeder Teresa miljoenen euro's van schenkers.
2. Alle Cyprioten zouden een hoge taks betalen op hun spaargeld.
3. Facebook onthoudt welke websites de gebruikers nog bezoeken.
4. De cursisten spraken op vrijdag 7 september met de politici.
5. Hij bezat de Europese titel sinds de zomer van 2014.
6. Dat is de belangrijkste rechtbank van het land.
7. Sterke lopers onder de veldrijders zagen hun kans.
8. De officiële resultaten zijn waarschijnlijk morgen, donderdag, bekend.
9. Arbeiders sloopten stukken van de tempel met bulldozers.
10. De Warmathon hoopt duizenden mensen op straat te krijgen.
11. De ziekenfondsen betalen sinds 2016 het remgeld terug voor kinderen.

12. De meeste pastoors zijn niet tevreden over aartsbisschop Léonard.
13. Vele tienduizenden mensen bekijken hun filmpjes op de website YouTube.
14. Met Pasen was minder dan één op vijf hotelkamers bezet.
15. De chefs bij Noma koken met producten uit de streek.

NEWS3

1. Dat vindt plaats op 25 september in Kopenhagen in Denemarken.
2. De pakjes brengt hij pas op 6 december.
3. Er zijn wedstrijden voor de best verklede bezoekers.
4. Op 6 december komt Sinterklaas langs.
5. Behalve in Brazilië, daar spreken mensen Portugees.
6. Voor de quizploeg probeert hij alles te onthouden.
7. De capsule in Boston zat er sinds 1914.
8. Eén straat heeft bijzondere parkeermeters.
9. Je hebt ook de sociale netwerken op internet, zoals Facebook.
10. Uiteindelijk bleken de toeschouwers toch in 'veilige' zones te staan.
11. Twee bedrijven uit Italië maken samen pasta.
12. Enkel president Obama kan de pijplijn nog tegenhouden.
13. In Groot-Brittannië vond het wereldkampioenschap darts plaats.
14. Na het wereldkampioenschap in Brazilië wilden ze snel naar huis.
15. Dit betekent net hetzelfde als keuze 2.

NEWS4

1. Die bleek 18 keer sterker dan eerst gedacht.
2. Haar tegenstanders blijven steken op 24 zetels.
3. Toen vond in Brazilië het wereldkampioenschap voetbal plaats.
4. De bibliotheek heeft 20 jaar lang cd's gekocht.
5. Moeder Teresa wordt op 4 september heilig verklaard.
6. Australië lijdt onder de zwaarste bosbranden sinds jaren.
7. Hij was 35 jaar sportjournalist op de radio.

8. In België is dat verboden op tijdelijke plaatsen.
9. Je kan dat tijdelijk gratis beluisteren op iTunes.
10. Dat zei de Amerikaanse president Obama op tv.
11. Tijdens het bezoek was er protest tegen Obama.
12. Dat jaar kwamen de eerste 600 bezoekers naar het park.
13. Elektronische maaltijdcheques kosten veel minder dan papieren cheques.
14. In 2013 stopt hij ook als president van China.
15. Dat was de zesde rally voor het Belgisch kampioenschap.

NEWS5

1. Op dit moment zijn er 800 strips beschikbaar.
2. Zondag kwamen de Europese ministers van Financiën samen in Luxemburg.
3. Belgische organisaties gebruikten 6 miljoen voor noodhulp.
4. Tanken langs de snelweg blijft heel duur.
5. Het decor drijft op het Bodenmeer.
6. Dat is dé auto in Oost-Duitsland.
7. De Turkse president Erdogan sprak het land toe.
8. De onderzoekers zetten nu aardbeiplantjes op duizend vensterbanken.
9. De spelers hadden achteraf kritiek op trainer Weiler.
10. Ook België heeft redders en dokters ter plaatse.
11. De eerste voorstelling vindt plaats op zondag 20 september.
12. De Britse zangeres Adele is met succes geopereerd.
13. De rechtbank bestaat sinds 2002.
14. Zodra de index 2 procent stijgt, helpt de overheid.
15. Bijvoorbeeld de presidenten van Rusland, China en Syrië.
16. Het gaat bijvoorbeeld om kwetsende opmerkingen op Facebook.

C PER results per participant

Table 5: Overview of the recognition errors made on the phoneme level for the consonants in percentages. **Blue bold** and **orange bold** numbers indicate the best and worst ASR performance per phoneme for both speaker groups. Numbers with blue and orange backgrounds represent the best and worst ASR performance per speaker. The final column gives the differences between the mean PERs of the oral cancer patients (M_{pt}) and the mean PERs of the healthy speakers (M_{hc}).

	Healthy							Patient								M_{pt} - M_{hc}
	01	**07**	**08**	**09**	**12**	*M*	*SD*	**02**	**03**	**04**	**05**	**06**	**11**	*M*	*SD*	
Plosive																
/p/	**2.4**	**5.4**	3.6	3.6	3	*3.6*	*1.1*	**57.5**	18.6	50.9	18.6	**13.8**	42.5	*33.7*	*18.9*	30.1
/b/	**0.6**	2.5	1.9	**8.9**	**0.6**	*2.9*	*3.5*	**62.7**	18.4	39.2	**5.7**	10.1	37.3	*28.9*	*21.6*	26
/t/	4.7	6.1	7.6	**9.1**	**3.9**	*6.3*	*2.1*	69	26.9	33.7	23.3	**16.2**	**77.5**	*41.1*	*25.7*	34.8
/d/	6.2	8.5	6	**12.4**	**5.5**	*7.7*	*2.9*	62.1	25.3	47.6	20.5	**14.7**	**78.2**	*41.4*	*25.4*	33.7
/k/	2	3.6	4.8	**7**	**0.4**	*3.6*	*2.5*	85.5	27.4	**97.2**	**11.3**	49.2	44.4	*52.5*	*33.1*	48.9
Nasal																
/m/	2.6	**2.2**	**2.2**	**5.7**	2.6	*3.1*	*1.5*	**57**	20.2	41.7	10.1	**9.2**	38.2	*29.4*	*19.3*	26.3
/n/	4.9	6.4	6	**7.3**	**2.8**	*5.5*	*1.7*	**46.1**	16.4	36.7	13.8	**8.7**	40.7	*27.1*	*15.9*	21.6
/ŋ/	**21.2**	15.2	19.7	10.6	**6.1**	*14.6*	*6.3*	89.4	48.5	**90.9**	**13.6**	50	57.6	*58.3*	*29*	43.7
Trill																
/r/	**3.2**	5.7	**6.6**	**6.6**	4.4	*5.3*	*1.5*	**61.5**	31.4	61.3	**11**	19	37.2	*36.9*	*21.1*	31.6
Fricative																
/f/	6.7	**2.2**	**13.3**	**13.3**	**2.2**	*7.5*	*5.6*	46.7	**22.2**	55.6	26.7	24.4	**71.1**	*41.1*	*19.9*	33.6
/v/	**3.8**	5	**3.8**	**6.9**	**3.8**	*4.7*	*1.4*	**58.5**	22.6	52.8	**10.7**	**10.7**	42.8	*33*	*21.2*	28.3
/s/	**2.6**	3.9	4.7	**5.7**	**2.6**	*3.9*	*1.4*	**63.8**	15.5	31.5	**10.3**	15	57.1	*32.2*	*23.1*	28.3
/z/	7.3	**5**	**9.5**	8.9	7.8	*7.7*	*1.7*	**70.9**	20.7	41.3	21.8	**20.1**	58.7	*38.9*	*21.9*	31.2
/x/	3	4.5	3	**5.3**	**1.5**	*3.5*	*1.5*	46.6	13.5	**51.1**	**12**	**12**	43.6	*29.8*	*19.1*	26.3
/ɣ/	6.7	8.3	**9.2**	7.5	**5**	*7.3*	*1.6*	61.7	23.3	**69.2**	**5**	13.3	68.3	*40.1*	*29.5*	32.8
/h/	14.9	18.9	**19.9**	14.9	**6.5**	*15*	*5.3*	41.8	17.4	36.3	**56.2**	**14.9**	52.7	*36.6*	*17.4*	21.6
Approximant																
/ʋ/	6.1	3.8	6.1	**9.9**	**5.3**	*6.2*	*2.3*	**58.8**	24.4	47.3	**6.9**	10.7	48.1	*32.7*	*21.7*	26.5
/l/	6.4	5.8	5.5	**8**	**4.2**	*6*	*1.4*	64.3	26	**68.5**	**11.6**	19.6	64.3	*42.4*	*26*	36.4
/j/	**7.6**	12.1	**18.2**	13.6	13.6	*13*	*3.8*	53	**25.8**	**71.2**	28.8	47	69.7	*49.3*	*19.4*	36.3

Table 6: Overview of the recognition errors made on the phoneme level for the vowels in percentages. **Blue bold** and **orange bold** numbers indicate the best and worst ASR performance per phoneme for both speaker groups. Numbers with blue and orange backgrounds represent the best and worst ASR performance per speaker. The final column gives the differences between the mean PERs of the oral cancer patients ($\boldsymbol{M}_{pt}$) and the mean PERs of the healthy speakers ($\boldsymbol{M}_{hc}$).

	Healthy							Patient								$\boldsymbol{M}_{pt}$ - $\boldsymbol{M}_{hc}$
	01	**07**	**08**	**09**	**12**	*M*	*SD*	**02**	**03**	**04**	**05**	**06**	**11**	*M*	*SD*	
Front																
/i(ː)/	**6.8**	12.3	11	**13.7**	9.6	*10.7*	*2.7*	71.2	29.5	**85.6**	**21.2**	43.2	55.5	*51*	*24.6*	40.3
/y/	**9.5**	14.3	14.3	14.3	**19**	*14.3*	*3.4*	**66.7**	19	38.1	**9.5**	19	57.1	*34.9*	*23.1*	20.6
/ɪ(ː)/	**3.9**	**4.6**	4.1	4.1	4.1	*4.2*	*0.3*	**40**	**17.6**	35.9	18.1	21.4	26.3	*26.6*	*9.5*	22.4
/e(ː)/	5	**7.1**	3.5	4.3	**1.4**	*4.3*	*2.1*	**58.9**	19.1	45.4	**12.8**	29.1	53.9	*36.5*	*19*	32.2
/ɛ/	**7.9**	10.8	**13.6**	13.3	**7.9**	*10.7*	*2.8*	**62**	28.7	40.9	**14.7**	24.4	44.4	*35.9*	*16.8*	25.2
Central																
/a(ː)/	**4.6**	6	6	**7.9**	6.5	*6.2*	*1.2*	**43.5**	16.2	30.1	**9.7**	17.1	35.2	*25.3*	*13*	19.1
/ə/	7.6	6.9	6.1	**9.4**	**4.3**	*6.9*	*1.9*	**51.5**	18	39.1	**15.8**	17.1	45.8	*31.2*	*16.1*	24.3
/ɵ/	4.5	6.8	**2.3**	**13.6**	4.5	*6.3*	*4.4*	**77.3**	31.8	59.1	25	**20.5**	52.3	*44.3*	*22.2*	38
Back																
/u/	**2.4**	7.1	**9.5**	7.1	**9.5**	*7.1*	*2.9*	**77.4**	29.8	59.5	**14.3**	21.4	53.6	*42.7*	*24.6*	35.6
/oː/	**3.5**	**3.5**	11	**14.5**	11.6	*8.8*	*5*	**68.6**	33.7	61	20.9	**16.3**	60.5	*43.5*	*22.7*	34.7
/ɔ(ː)/	3.3	3.7	5.1	**6.5**	**2.3**	*4.2*	*1.6*	**55.3**	37.7	54	31.2	**10.7**	45.6	*39.1*	*16.7*	34.9
/ɑ/	3.9	5.5	3.2	**8.1**	**2.3**	*4.6*	*2.3*	**48.1**	30.6	40.3	13.9	**8.4**	40.3	*30.3*	*15.9*	25.7
Diphthongs																
/ɛɪ/	3.9	**2.6**	5.2	**11.8**	3.3	*5.4*	*3.7*	42.5	**9.2**	39.9	9.8	14.4	**54.9**	*28.5*	*19.7*	23.1
/ʌʊ/	**12.9**	**3.2**	**3.2**	**12.9**	**3.2**	*7.1*	*5.3*	32.3	16.1	29	16.1	**12.9**	**54.8**	*26.9*	*15.8*	19.8
/œy/	**11.4**	**0**	6.8	**0**	**0**	*3.6*	*5.2*	**50**	15.9	36.4	**4.5**	**4.5**	38.6	*25*	*19.3*	21.4