Making speech technology accessible for pathological speakers

Bence Márk Halpern

The research leading to this dissertation has received funding from the European Union's Horizon 2020 research and innovation programme under MSC grant agreement No 766287. The Department of Head and Neck Oncology and Surgery of the Netherlands Cancer Institute receives a research grant from Atos Medical (Horby, Sweden), which contributes to the existing infrastructure for quality of life research.

Published by: Ridderprint | www.ridderprint.nl
Cover design: Sandra Tukker
ISBN: 978-94-6458-494-3

Making speech technology accessible for pathological speakers

ACADEMISCH PROEFSCHRIFT

ter verkrijging van de graad van doctor
aan de Universiteit van Amsterdam
op gezag van de Rector Magnificus
prof. dr. ir. P.P.C.C. Verbeek
ten overstaan van een door het College voor Promoties ingestelde commissie,
in het openbaar te verdedigen in de Agnietenkapel
op maandag 3 oktober 2022, te 14.00 uur
door Bence Márk Halpern
geboren te Budapest 13

Promotiecommissie

Promotor:	prof. dr. M.W.M. van den Brekel	Universiteit van Amsterdam
Copromotores:	dr. R.J.J.H. van Son	Universiteit van Amsterdam
	dr. O.E. Scharenborg	Technische Universiteit Delft
Overige leden:	prof. dr. L.E. Smeele	Universiteit van Amsterdam
	prof. dr. P.P.G. Boersma	Universiteit van Amsterdam
	prof. dr. A.P.J. van den Bosch	Universiteit van Amsterdam
	prof. dr. T. Toda	Nagoya University
	prof. dr. G. Van Nuffelen	Universiteit Gent
	dr. J.A. Burgoyne	Universiteit van Amsterdam
	dr. B. Sisman	National University of Singapore

Faculteit der Geesteswetenschappen

Contents

NOTATIONS AND SHORTHANDS

The following shorthands and notations will be used throughout the thesis.

- **AM** - Acoustic Model
- **ASR** - Automatic Speech Recognition
- **CER** - Character Error Rate
- **DNN** - Deep Neural Network
- **DTW** - Dynamic Time Warping
- **E2E** - End-to-End
- **EM** - Expectation Maximisation
- **FHVAE** - Factorised Hierarchical Variational Autoencoder
- **fMLLR** - Feature space Maximum Likelihood Regression
- **GT** - Ground Truth
- **GV** - Global Variance
- **HMM** - Hidden Markov Model
- **LASSO** - Least Absolute Shrinkage and Selection Operator
- **LM** - Language Model
- **LR** - Learning Rate
- **LTAS** - Long Time Average Spectrum
- **MCD** - Mel Cepstral Distortion
- **MFCC** - Mel Frequency Cepstral Coefficients
- **MoA** - Manner of Articulation
- **MOS** - Mean Opinion Score
- **MS** - Modulation Spectrum
- **N2D** - Normal-To-Dysathric
- **PER** - Phoneme Error Rate

- **PoA** - Place of Articulation
- **PSOLA** - Pitch Synchronous Overlap Add
- **RNN** - Recurrent Neural Network
- **RQ** - Research Question
- **SotA** - State of the Art
- **SLP** - Speech Language Pathologist
- **STOI** - Short Time Objective Intelligibility
- **SNR** - Signal to Noise Ratio
- **TTS** - Text-To-Speech (Synthesis)
- **VAE** - Variational Autoencoder
- **VC** - Voice Conversion
- **VQVAE** - Vector Quantised Variational Autoencoder
- **WER** - Word Error Rate
- /a/ - Indicates a phoneme. In Chapter 2, this notation is reserved for IPA, while in Chapter 3 we use the ARPAbet
- Vectors will be usually denoted with lowercase boldface (**a**), while matrices will be denoted with uppercase boldface (**A**)
- x_i or $x(i)$ - the ith element of the vector. In Chapter 4, $x(i)$ will be used for element, when the subscript is used to make a named distinction (e.g. x_p being pathological x)
- $\mathscr{L}$ - Loss function
- $\mathcal{N}(0, \mathbf{I})$ - a normal (Gaussian) distribution with zero mean and unit variance
- $\hat{x}$ - estimator of x (e.g. an estimate of speech severity)
- $\mathbb{E}$ - the expected value of a random variable
- $\mathbb{I}$ - the indicator function
- $\mathbb{R}$ - the set of real numbers
- $\mathbb{R}^+$ - the set of positive real numbers
- $\mathbb{R}^d$ - a real-valued vector of length d
- $\mathbb{Z}$ - the set of whole numbers

- $A \in B$ - A is element of a set B
- $|\cdot|$ - the cardinality of a set. The absolute value will be denoted using the 1-norm to avoid notation overload
- $||\cdot||_p$ - the p-norm
- $KL(\cdot||\cdot)$ - the Kullback-Leibler divergence between two probability distributions
- p - Pearson's correlation
- ρ - Spearman's correlation

SUMMARY

Making speech technology accessible for pathological speakers

Speech technology has become widespread in the past decades: voice assistants, voice biometrics, automated call centres, autotune, and vocaloid idols are part of our everyday lives. The rapid proliferation of these technologies were enabled by the new field of deep learning, which allowed large scale pattern recognition using previously unseen amounts of data. These technologies are beneficial and convenient for the general population, but certain parts of the population, e.g., atypical and pathological speakers, these technologies do not work well. It is especially problematic that pathological speakers are less able to use these technologies as they are often physically disabled, meaning they would have a strong need for voice assistants. Apart from voice assistants, there are several other kinds of speech technologies where the main user would be a pathological speaker. However, these speech technologies currently do not work well for these pathological speakers. In other words, speech technology lacks accessibility to pathological speakers. This thesis presents a series of studies towards making pathological speech accessible to pathological speakers. These studies concern three applications of speech technology: automatic speech recognition for atypical and pathological speech, automatic speech severity evaluation for oral cancer speakers, and pathological voice conversion.

Chapter 2: Towards Inclusive Automatic Speech Recognition

Chapter 2 investigates how accessible is automatic speech recognition (ASR) for atypical speakers of Dutch and Mandarin. To investigate this, we define the term bias metric, which we calculate by subtracting the lowest (=best) WER from the WER of each atypical speaker group investigated. The study finds significant bias against (Dutch) male speech, children's speech, old adults' speech, and non-native (Dutch and Mandarin) accented speech. We argue that the training data, pronunciation, type of ASR architecture and language can all be a source of bias.

Chapter 3: Low-Resource Automatic Speech Recognition and Error Analyses of Oral Cancer Speech

Oral cancer patients often receive chemoradiation during their treamtent. As a side-effect of that treatment, oral cancer speakers become temporarily weak and limited in their moving, meaning that they need a lot of help in this period. Patients could use voice assistants to assist them during these difficult times, however, automatic speech recognition currently does not work well for oral cancer speakers. To address this issue, **Chapter 3** develops and compares different systems for the recognition of oral cancer speech. As part of our study, we propose a speaker adaptation-based approach for

recognising oral cancer speech, attaining 7.7% absolute improvement over our baseline trained on healthy speech. Furthermore, the study finds that plosives and some vowels (/aa/ and /uw/) were challenging to recognise for the developed ASR systems, which are known to be impacted in the case of oral cancer speakers.

Chapter 4: Automatic Evaluation of Spontaneous Oral Cancer Speech Using Ratings from Naive Listeners

Many oral cancer patients have impaired speech due to the treatment of oral cancer. To address their speech impairment, oral cancer patients often require speech therapy. We often want to measure the efficiency of speech therapy, meaning that we need a way to track the current level of speech impairment during speech therapy. Estimating speech impairment is currently done by speech-language pathologists, however, this estimation have several shortcomings. **Chapter 4** compared the ASR systems developed in **Chapter 3** with acoustic feature-, and comparison-based methods to assess whether it is possible to predict speech severity ratings automatically from spontaneous oral cancer speech samples, and which method is the best for predicting this. The best techniques using explainable regression models in the study correlate highly with an expert listener's severity ratings. We simultaneously collect severity ratings from non-expert listeners and show that their ratings correlate highly with the expert listener.

Chapter 5: Pathological voice adaptation with autoencoder-based voice conversion

Patients waiting for oral cancer surgery experience anxiety, which is exacerbated by uncertainty about what will happen to them after the surgery. It would be important to alleviate patients' anxiety, as studies show that this anxiety has a negative impact on their quality of life, even after the surgery. One source of uncertainty is regarding the severity of the speech impairments patients might have after surgery. It would be important to provide patients with more information about their future speech impairment. One possible way to inform them would be showing them synthesised speech samples of their future speech impairment. Therefore, it would be important to procure a tool that can show how patients might sound after oral cancer treatment. A possible tool would be based on voice conversion, which is the main topic of **Chapters 5** and **6**. **Chapter 5** proposes a voice conversion setup which converts pathological speech to a speech of another pathological speaker. The proposed voice conversion framework produces speech samples perceptually similar to the target pathological speaker's voice characteristics. to the target speaker in the case of low and high severity while also demonstrating reasonable naturalness. However, the framework does not allow the synthesis of arbitrary utterances, which is a limitation.

Chapter 6: Towards Identity Preserving Normal to Dysarthric Voice Conversion

To address this limitation, **Chapter 6** proposes a two-step voice conversion approach. In the first step, a sequence to sequence (seq2seq) model is used to convert healthy speech to pathological speech. The first step allows synthesis of arbitrary utterances, however the healthy speaker's identity is not retained during the conversion. For this

purpose, we employ a second step which is targeted to regain this identity based on the model proposed in **Chapter 5**. The proposed two-step approach controls speech severity according to three objective speech severity measures. It also achieves reasonable naturalness, however, it lacks similarity to the healthy speakers.

Discussion and concluding remarks

The studies in Chapters 2 through 6 also point out that standard evaluation measures in speech technology, such as the word error rate to evaluate automatic speech recognisers and the mean opinion score to evaluate the naturalness of generated speech, have several shortcomings. The word error rate does not explicitly capture the performance of automatic speech recognisers on underrepresented groups in the test data. We suggest including more diverse speaker groups in automatic speech recognition testing and quantifying the bias as a new standard for evaluation. The mean opinion score is not only sensitive to naturalness but also to other factors such as speech severity, age, gender, and openness of the raters toward speech technology applications. We suggest that objective naturalness measures should be developed to alleviate the shortcomings of the mean opinion score. In general, we urge earlier testing with broader users of speech technology to facilitate accessibility of speech technology for all users, including pathological speakers.

SAMENVATTING

Spraaktechnologie is de afgelopen decennia snel ingeburgerd: Spraakassistenten, spraakbiometrie, geautomatiseerde callcenters, autotune en Vocaloïde idolen maken nu deel uit van ons dagelijkse leven. De snelle groei van deze technologieën is mogelijk gemaakt door doorbraken op het gebied van machinaal leren met neurale netwerken, het zogenaamde "Deep Learning". Deze doorbraken hebben patroonherkenning op grote schaal mogelijk gemaakt met behulp van ongekende hoeveelheden data. Deze spraaktechnologieën zijn nuttig voor iedereen, maar ze werken minder goed, of zelf helemaal niet, voor bepaalde groepen in de samenleving, zoals atypische en pathologische sprekers. Het is met name wrang dat mensen met een spraakbeperking deze technologieën minder of helemaal niet kunnen gebruiken omdat zij vaak ook fysieke beperkingen hebben waardoor juist zij baat zouden hebben bij, b.v, spraakassistenten. Naast spraakassistenten zijn er verschillende andere soorten spraaktechnologieën waarbij de belangrijkste potentiële gebruikers een spraakbeperking hebben. Dit proefschrift bevat een reeks studies die tot doel hebben om spraaktechnologie toegankelijk te maken voor sprekers met een spraakbeperking. De studies hebben betrekking op drie toepassingen van spraaktechnologie: automatische spraakherkenning voor atypische en pathologische spraak, automatische evaluatie van de ernst van de spraakbeperking voor patiënten die behandeld zijn voor mondholtekanker en pathologische spraakconversie.

Hoofdstuk 2: Op weg naar inclusieve automatische spraakherkenning

Hoofdstuk 2 onderzoekt hoe toegankelijk de automatische spraakherkenning (ASR) is voor atypische sprekers van het Nederlands en het Mandarijn Chinees. Om dit te onderzoeken, introduceren we de term bias-metriek, die we berekenen als de laagste (= beste) foutscore (WER) van elke atypische groep sprekers. De studie vindt een aanzienlijke verslechtering bij de herkenning van geaccentueerde spraak van mannelijk sprekers (Nederlands), spraak van kinderen, de spraak van oudere volwassenen en niet-moedertaal sprekers (Nederlands en Mandarijn). Wij concluderen dat het trainingsmateriaal, de uitspraak, de architectuur van de spraakherkenner en de taal allen een bron van bias kunnen zijn bij de herkenning.

Hoofdstuk 3: Automatische spraakherkenning en foutanalyses met weinig spraakdata van spraak na behandeling voor mondholtekanker

Patiënten met mondholtekanker krijgen vaak chemo-radiotherapie tijdens hun behandeling. Als een bijwerking van die behandeling kunnen patiënten tijdelijk aan huis gebonden zijn met een beperkte mobiliteit. Deze patiënten zouden gedurende deze moeilijke periode geholpen kunnen zijn met spraakassistenten. Echter, automatische spraakherkenning werkt momenteel niet goed voor sprekers die behandeld zijn voor mondholtekanker. **Hoofdstuk 3** ontwikkelt en vergelijkt verschillende systemen voor de

herkenning van de spraak van patiënten die behandeld zijn voor mondholtekanker. Als onderdeel van onze studie presenteren wij een spreker-adaptief gebaseerd systeem voor de herkenning van deze spraak dat een absolute verbetering bereikt van 7,7% boven het referentiesysteem getraind op spraak van gezonde sprekers. Uit de studie blijkt ook dat de herkenning van plosieven en sommige klinkers ("aa" en "oe"), waarvan bekend is dat ze beïnvloed worden door behandeling voor mondholtekanker, problemen opleverden voor de ontwikkelde spraakherkenners.

Hoofdstuk 4: Automatische evaluatie van spontane spraak na mondholtekanker met behulp van beoordelingen van naïeve luisteraars

Na een behandeling voor mondholtekanker krijgen veel patiënten problemen met spreken. Voor deze spraakstoornissen krijgen patiënten vaak spraaktherapie aangeboden. Om de effectiviteit van de logopedie te bepalen is een manier nodig om het verloop van de spraakstoornis te volgen tijdens de therapie. De ernst van de spraakstoornis wordt nu nog beoordeeld door de logopedist, maar een dergelijke subjectieve beoordeling heeft nadelen. In **Hoofstuk 4** worden de spraakherkenningssystemen van **Hoofdstuk 3** vergeleken met methoden gebaseerd op akoestisch features en systeem-vergelijkingen om te onderzoeken of het mogelijk is om de beoordelingen van de logopedisten automatisch te voorspellen op grond van opnamen van spontane spraak van patiënten na behandeling voor mondholtekanker, en welke methode het beste resultaat geeft. De beste technieken in de studie die gebruik maakten van verklaarbare regressiemodellen correleren sterk met de beoordelingen van de expert. Tegelijkertijd verzamelden we beoordelingen van de ernst van de spraakstoornis door naïeve luisteraars en laten we zien dat hun beoordelingen sterk correleren met die van de expert.

Hoofdstuk 5: Pathologische spraakaanpassing met spraakconversie op basis van een autoencoder

Patiënten die wachten op een operatie voor mondholtekanker hebben vaak zorgen en angst voor de toekomst, die worden verergerd door onzekerheid over hoe hun leven er na de operatie uit zal zien. Het is belangrijk om de zorgen van patiënten te verminderen, omdat studies aantonen dat onzekerheid en verkeerde verwachtingen een negatieve invloed hebben op de kwaliteit van leven, zelfs na de operatie. Een bron van onzekerheid voor patiënten zijn de spraakstoornissen die kunnen optreden na hun operatie en de ernst van de stoornissen. Het is belangrijk om patiënten meer informatie te kunnen geven over hun toekomstige spraak en de verwachte problemen daarmee. Een mogelijke manier om hen beter te informeren zou zijn om hen computer gegenereerde spraak van hun mogelijke toekomstige spraakstoornissen te laten horen. Hiervoor zou het nuttig zijn om een hulpmiddel te ontwikkelen dat kan laten horen hoe patiënten kunnen gaan klinken na behandeling voor mondholtekanker. Een mogelijke manier om dit te bereiken zou gebaseerd kunnen worden op spraakconversie, het belangrijkste onderwerp van **Hoofstukken 5 en 6**. **Hoofdstuk 5** stelt een systeem voor stemconversie voor die pathologische spraak omzet naar de spraak die klinkt als gesproken door een andere pathologische spreker. Het voorgestelde conversie framework produceert voorbeelden van pathologische spraak die klinken als de doelspreker, maar dan met meer of minder

ernstige spraakstoornissen. Dit systeem is echter niet in staat om willekeurige zinnen te produceren wat de toepassingen beperkt.

Hoofdstuk 6: Naar behoud van spreker identiteit bij een stemconversie van normale naar dysartrische spraak

Om deze beperking aan te pakken, stelt **Hoofstuk 6** een tweetraps stemconversie-aanpak voor. In de eerste stap wordt een sequentie-naar-sequentie (seq2seq) model gebruikt om gezonde spraak om te zetten in pathologische spraak. De eerste stap maakt synthese van willekeurige uitingen mogelijk, maar de identiteit van de gezonde spreker blijft niet behouden tijdens de conversie. Voor dit doel gebruiken we een tweede stap die is gericht op het weer converteren naar de oorspronkelijke identiteit op basis van het model dat wordt voorgesteld in **Hoofstuk 5**. De voorgestelde tweetraps benadering controleert de ernst van de spraakafwijking volgens drie objectieve meetmethoden. Er wordt een redelijke mate van natuurlijkheid bereikt, maar de huidige techniek levert nog onvoldoende gelijkenis met de oorspronkelijke spraak.

Discussie en slotopmerkingen

Het onderzoek beschreven in de **Hoofstukken 2** tot en met **6** wijzen er ook op dat standaard meetmethoden in spraaktechnologie, zoals de woord foutscore om automatische spraakherkenners te evalueren en de gemiddelde opiniescore om de natuurlijkheid van gegenereerde spraak te evalueren, een aantal tekortkomingen hebben. De woord foutscore geeft niet expliciet de prestatie weer van automatische spraakherkenners bij ondervertegenwoordigde groepen in de resultaten. Wij raden aan om meer diverse sprekersgroepen op te nemen in automatische spraakherkenningstests en het kwantificeren van de bias te gebruiken als een nieuwe evaluatiestandaard. De gemiddelde opiniescore is niet alleen gevoelig voor natuurlijkheid, maar ook voor andere factoren zoals de ernst van de spraakafwijkingen, leeftijd, geslacht en de mate waarin beoordelaars openstaan voor toepassingen van spraaktechnologie. Wij stellen voor om objectieve natuurlijkheidsmaten te ontwikkelen om de tekortkomingen van de gemiddelde opiniescore van luisteraars te verlichten. Over het algemeen dringen we aan op eerder testen met een diversere groep gebruikers van spraaktechnologie om de toegankelijkheid van spraaktechnologie voor alle gebruikers te vergemakkelijken, inclusief pathologische sprekers.

1

INTRODUCTION

1.1. ACCESSIBILITY OF SPEECH TECHNOLOGY

Speech technology has become ubiquitous in the past decade. Voice assistants are now built into our mobile phones and computers. Call centres are increasingly replacing the human workforce with (semi-)automated conversational agents. The entertainment media and music industry employs automated pitch manipulation techniques ("auto-tune") on a daily basis. Traditional idols are increasingly challenged by vocaloid idols, who are the new stars of Asian entertainment media. Speaker recognition is used as voice biometrics, and online webinars are transcribed by speech recognisers on a massive scale. A great deal of development happened in this area, in a seemingly very short time span.

This proliferation of speech technology is due to the advent of deep learning and related advanced techniques enabled by graphical processing units (GPU). Deep learning simultaneously enables automation of pattern recognition and using dataset sizes that were previously unconceivable. The ability to use massive amounts of data for certain tasks led to improved performance on many tasks, including the ImageNet image classification benchmark [1]. This benchmark is arguably one of the most influential ones in deep learning, as architectures successful on this benchmark often turned out to be successful on a variety of other tasks.

However, as often happens in history, improvements come with considerable societal costs. These societal problems include the carbon footprint of using GPUs [2], but also racial/ethnical bias creeping into these models [3]. The source of these biases is often unclear, but often it is the composition of the dataset used to train the models [4]. Some people argue this bias happens because deep learning models work as "stochastic parrots", effectively spitting back to us all the societal biases that are present in the data [5].

Speech technology is no exception to these biases. It has, for instance, been well known for a while that automatic speech recognisers (the transcription of speech to text) are not accessible to many people, e.g. speakers with an accent, dialect [6], or speech

pathology [7]. Apart from scientific evidence, anecdotal evidence also seems to corroborate these findings. Most likely everyone has an experience concerning the fragility of these technologies, e.g., Siri misunderstanding a word due to an unclear pronunciation, or Cortana refusing to react to the wakeword for the fifth time.

1.2. TERMINOLOGICAL COMMENTS

Before diving into the technological difficulties and issues of accesible speech technology, we need to achieve common ground on some speech terms used throughout this thesis. Reaching this common ground is especially important with regards to the definition of intelligibility, as there is a reported lack of consensus on the usage of the term [8, 9].

We define speech technology first. Speech technology is defined here as a blanket term for any kind of computer-based technology that uses human speech. The technology can be either a complete business solution, or a specific algorithm. Examples of speech technology will be given in Section 1.4.

The definition of pathological speech is in itself not easy. The American Speech-Language Hearing Association uses pathological speech as an umbrella term for (1) speech disorders, (2) language disorders, (3) social communication disorders, (4) cognitive communication disorders and (5) swallowing disorders [10]. We will look at two speech disorders in this thesis: dysarthric and oral cancer speech. “Pathological speech” here is an umbrella term for these speech disorders.

Apart from pathological speech, we will often mention the term atypical speech. Atypical speech is a broader term than pathological speech, as it includes any type of speech that is deviant from typical adults in one particular language group. Examples include accented speech, speech with dialect, old speech or child speech, and also pathological speech.

When the expression ’pathological speech’ is used in this thesis, the interest is limited to two properties of speech. The first property is the intelligibility of the speech, which is defined as the extent that the orthography of the speech can be transcribed by other listeners solely based on acoustic cues, following [11]. However, the definition should be interpreted more liberally than in the context given by [11] as "other listeners" can also mean machine listeners, i.e., automatic speech recognisers. We think intelligibility is the most important aspect of pathological speech to be measured - if other people do not understand the pathological speakers then the speakers will feel frustrated, which will influence their quality of life. This property is expressed quantitatively, e.g., by a percentage of unrecognised words/characters/phonemes by a listener.

The second property is the severity of the speech. Speech severity is a somewhat non-ideal term use as it is easily mistaken for the severity of the underlying disease. For this reason, often similar concepts, such as voice quality, acceptability, and perceived healthiness are used to described a similar phenomenon [12, 13]. Still, severity is widely adopted in the speech technology literature, therefore, we will use this term for the sake of consistency [14, 15, 16, 17]. Following the definition of acceptability by [13], severity is defined by the degree to which speech calls attention to itself apart from the content. In other words, the degree that speech deviates from typical. Severity is a distinct but related phenomenon to intelligibility, for example, high severity speakers are almost

always low intelligibility speakers. However, there are often highly intelligible speakers with fairly unusual voice characteristics, e.g., with a breathy, creaky or hoarse voice. Severity is also meant to be an encapsulating term for all these acoustic anomalies in the speech.

While it is not a speech pathology term, it is worth mentioning the term "naturalness", as it can be easily confused with severity. In this thesis naturalness, refers to the fidelity (voice quality) of the computer synthesised speech. Putting it in a different way, it is a measure of the extent that computer synthesised speech is indistinguishable from real speech. Naturalness and its evaluation will be further discussed in Section 1.4.3.

1.3. USE CASES OF ACCESSIBLE SPEECH TECHNOLOGY

In the present thesis, we will look at a selection of closely intertwined use cases of speech technology for pathological and atypical speakers: whether these techniques break down or not, how these techniques break down, and what we can do to address these issues.

1.3.1. DYSARTHRIC SPEECH RECOGNITION

The irony of fate is that speech technology does not work with users who would benefit the most from it. One such example are dysarthric speakers. Dysarthria is 'a collective name for a group of speech disorders resulting from disturbances in muscular control over the speech mechanism due to the damage of the central or peripheral nervous system' [18]. The total number of people suffering from dysarthria are difficult to estimate, but more than 1 million people [19] are affected just by Parkinson's Disease in the US, 90% of which suffer from dysarthria sooner or later during the course of the disease [20].

A large portion of dysarthric speakers are often unable to move their limbs, therefore they are unable to use keyboards. Automatic speech recognition (ASR) could potentially be used as a replacement for keyboard-based input. The potential positive effect of ASR as an alternative input method has already been demonstrated in a user satisfaction study [21]. However, the performance of ASR for dysarthric users is currently lacking [22, 23]. Not being able to turn the lights on using a voice assistant is a minor inconvenience for most people but more than frustrating for a person with dysarthria who has trouble moving without assistance.

There are many avenues to improve dysarthric speech recognition. One of the avenues attracting considerable attention recently is data augmentation. Data augmentation attempts to improve the performance of data-driven machine learning models by generating more data. In this thesis, we are going to look at a particular data augmentation technique called voice conversion. Voice conversion could be used to create artificially generated dysarthric speech which can be used as additional training material for the ASR and subsequently potentially improve dysarthric speech recognition.

1.3.2. ORAL CANCER SPEECH TECHNOLOGY

Oral cancer patients are in grave need of several different kinds of speech technology. First, these patients often receive surgery and chemoradiation during their treatment. Especially the latter can make patients weak and often temporarily physically disabled. During this period, patients could be supported by voice assistants. However, we will

see in Chapter 3 that oral cancer patients are not understood very well by ASRs, which means that patients are unable to rely on these voice assistants.

Second, our clinical experiences and findings in the literature show that patients are unsatisfied by the amount and quality of information received about the treatment, especially regarding the side-effects and handicaps caused by the treatment [24]. Several studies further pointed out that this is an important problem as insufficient counselling has a long-lasting effect on patients' quality of life [24, 25]. Therefore, it would be imperative to procure a tool that is able to show an example of the speech-related side effects to the patients. This problem could be potentially addressed using a technique called voice conversion, which we will further explain in Section 1.4.3.

Third, after oral cancer treatment, patients often need speech therapy. Currently, the success of speech therapy is evaluated using subjective tools, such as the Grade-Roughness-Breathiness-Astenicity-Strain Scale (GRBAS) and the Consensus Auditory Perceptual Evaluation of Voice (CAPE-V) [26]. These subjective evaluations are expensive and suffer from reliability issues [27]. Therefore, it would be important to replace these approaches with more reliable, objective measures [28, 29, 30]. However, objective measures also have their own set of shortcomings. One is that these models are often not not transparent in their decision-making process, i.e., they are suffering from a lack of *explainability*. The other problem is that these objective measures are using read speech instead of spontaneous speech, which might be not reflective of patient's real communicative problems. We will further explain automatic severity evaluation in Section 1.4.2.

1.4. SPEECH TECHNOLOGY TASKS

As mentioned above, we will focus on a limited subset of speech technology applications, namely: automatic speech recognition, automatic severity estimation, and voice conversion. We will briefly detail these tasks in this section, and we will present the current issues surrounding their direct application to the above-mentioned pathological speech use cases.

1.4.1. AUTOMATIC SPEECH RECOGNITION

The goal of automatic speech recognition (ASR) is to transcribe spoken utterances of speakers solely based on the acoustic speech signal. ASR is also often called "speech-to-text" (STT) which is a term inspired by the name of the key input and output of the task. This technology is used in several application contexts. It is used as the key component of automatic transcription services, but also in voice assistants such as Siri or Alexa.

Figure 1.1 shows the main parts of a traditional ASR system. First, the recorded speech is sampled and quantised into a digital signal. This digital signal is often called the raw or unprocessed waveform to contrast with the processed versions of the waveform. After this step, acoustic features are extracted, which are processed representations of the original raw waveform. Using these features, the acoustic model (AM) estimates the probability of a phoneme sequence given the acoustic features. In other words, the AM is the part of the ASR that connects the acoustics to the phonemes. The language model (LM) uses the lexicon to produce the probability of a certain word sequence. Finally, the decoder selects the most likely word sequence given the constraints

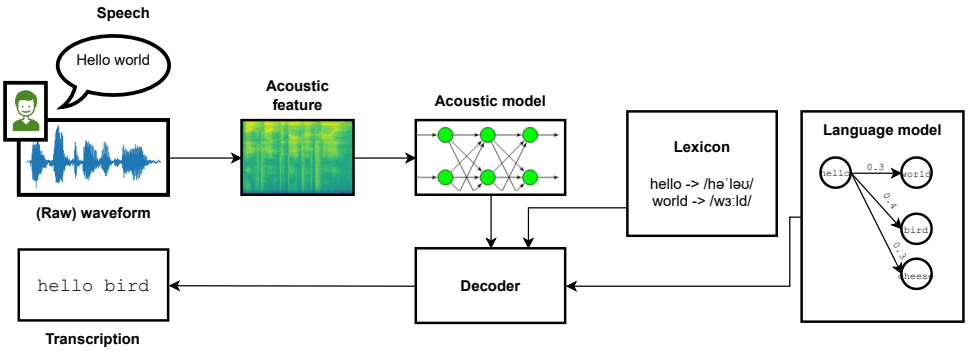


Figure 1.1: Main structural components of a traditional automatic speech recognition (ASR) system showing the steps from the speech to the transcription (text). Please note that the figure is greatly simplified.

imposed by the LM and the AM.

Newer, so-called end-to-end approaches tend to reduce this complexity and directly predict the word sequences from acoustic features, or even from the raw waveform. The goal of automatic speech recognition research is to improve the performance of speech recognisers by improving the individual components of the process illustrated on Figure 1.1. For example, in Chapter 3, we will focus on the selection of the best acoustic feature for oral cancer speech recognition. This selection process is often called feature engineering.

ASR tasks are most often evaluated using the word error rate or character error rate measure. The word error rate (WER) is defined as the sum of substitution (S), insertion (I), and deletion (D) errors in reference to the ground truth transcription divided by the total number of words (N) and multiplied by 100,

$$\text{WER} = \frac{S + I + D}{N} \cdot 100\%. \tag{1.1}$$

Word error rates are high in the case of pathological and atypical speech. There are multiple possible reasons for the current high WERs, but in general, it is still unclear what causes these high WERs for pathological and atypical speech compared to typical speech. An important reason is the low amount of data available for training pathological speech recognisers, as the AM of ASRs require a large amount of training data. The other reason is that most linguistic resources (e.g. pronunciation lexicons) are based on standard dialects and speech, and they are costly to adapt to different languages.

Moreover, potentially, pathological speech is just inherently more difficult (or sometimes impossible) to recognise, as even human listeners seem to struggle with pathological speakers. However, we do not know whether it is the same set of sounds that cause problems to ASRs and human listeners. To give an example, we have clinical evidence that the production of /p/ is difficult for oral cancer speakers [31, 32], but we do not know if it is words with /p/ that causes problems for speech recognisers.

1.4.2. AUTOMATIC SEVERITY ESTIMATION

Automatic severity estimation aims to estimate human severity scores based on the characteristics of the speech signal. The technology is increasingly being adopted by clinicians (typically, speech language pathologists) who want to monitor the progress of speech therapy [33].

Currently, severity estimation is done by speech-language pathologists (SLPs) using standardised subjective assessment tools, examples include the Consensus Auditory Perceptual Evaluation of Voice (CAPE-V) [26] or the Graded Roughness Breathiness Aesthenia Scale (GRBAS) [27]. These methods usually produce one or multiple scores for the speaker, these are called subjective scores. The aim of automatic severity estimation tools is to estimate the subjective scores of the SLPs with high accuracy. A higher correlation between the predicted (objective) scores and the subjective scores indicates a better performance of the automatic severity estimation tool.

There are different automatic severity estimation tools. In Chapter 3, we classify severity evaluation techniques into three categories: (1) techniques that use a combination of acoustic features and a statistical model to predict an objective severity score directly, (2) techniques that use automatic speech recognition errors as a proxy for severity evaluation, (3) techniques that use an error function to compare the reference (healthy) and the pathological speech's representation as a proxy for severity evaluation.

A lot of automatic severity estimation tools lacks transparency in their decision making process. When severity scores are directly predicted from the speech, it remains uncertain for the users what acoustic cues contributed most to the decision. This uncertainty is problematic because clinicians want to make sure that the decision is based on pathology-specific evidence (e.g. breathiness), rather than sociolinguistic, and extralinguistic cues (e.g. accent).

When severity scores are predicted from the word error rate, the decision making process is slightly more transparent, as the user can directly see what phonemes were misunderstood. However, there is no guarantee that the actual phoneme errors are due to the problems with the severity of the speech. It could very well be that there are some ASRs that are bad at recognising certain phonemes, e.g. phonemes that share their manner of articulation.

To summarise, there would be a need for automatic severity estimation technology which can *explain* what acoustic cues are important for the decision, and produce errors that are justifiable.

Another challenge is the sensitivity of severity evaluation to the context of the communication. Severity evaluation models are typically trained and tested with read speech. Read speech is not neccesarily representative of the real communicative environment and needs of the patients. For example, in the case of oral cancer speech, [34] concluded that in spontaneous oral cancer speech, plosives do not seem to be impacted, contrary to the result of many previous clinical studies on read speech [31, 32]. Therefore, there is a need for more *ecologically valid* severity evaluation tools: tools that are representative of patients' communicative difficulties in real conversations.

1.4.3. VOICE CONVERSION

Voice conversion (VC) aims to change someone's voice so that it sounds like someone else's voice. A typical voice conversion pipeline takes in a single utterance from a so-called "source speaker" and converts it to the characteristics of a different, "target speaker", as illustrated in Figure 1.2. In VC, the linguistic content of the speech remains intact (see bubble), but the waveform changes to reflect the change in their vocal characteristics and speaking style (see the **bold** typeface, indicating stress).

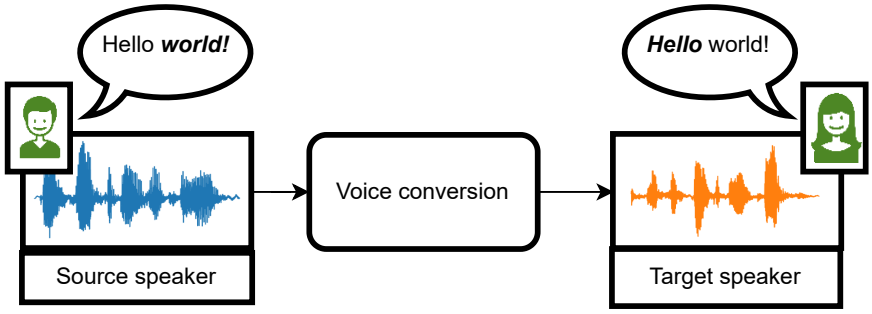


Figure 1.2: A voice conversion setup with a source speaker identifying as male and a target speaker identifying as female.

Voice conversion can be used for several applications. Voice anonymisation is an example application, which allows to hide the identity of the speakers, e.g. when collecting personal speech data in a GDPR-compliant way [35]. Enhancement of speech intelligibility is another example, where VC has been already used to convert unintelligible pathological voices (e.g. electrolaryngeal voice [36], dysarthric speech [37]) to more intelligible speech.

A more questionable application of VC is voice impersonation attacks. Such attacks are often used to get unauthorised access to bank accounts, or to cause political turmoil as recently demonstrated during the Russia-Ukraine war [38]. These artificial speech samples are often called "deepfakes" or "spoofs", and they pose a difficult moral question to researchers. To address the ethical aspects of voice conversion development, the ASVSpoof initiative was launched in 2017 to develop countermeasures against spoofed speech [39].

In general, VC applications assume that speech can be broken down into two distinct components, a component pertaining to the speaker identity and another component related to the linguistic content. The speaker identity component is sometimes called the time-invariant or supersegmental component (or level), while the linguistic component is called the time-variant or suprasegmental component (or level) of the speech signal. With pathological speech, the key difficulty is that the *disentanglement* of these two components is not straightforward: the speech pathology affects both the time-invariant and the time-variant levels of the speech signal.

We can partition VC methods into two distinct categories based on the type of data they need from the source and target speakers. These two categories are called parallel and non-parallel. Parallel VC requires data from the source and target speaker with the

same linguistic contents (i.e., same spoken words). Non-parallel VC still requires data from both speakers, but the linguistic content is different in that case. Parallel VC usually performs better than non-parallel VC because only the speaker characteristics have to be changed during the training. However, non-parallel allows more flexibility, which is often preferred in clinical applications, where obtaining parallel recordings can be tedious.

Voice conversion systems need to be evaluated to check if (1) the desired naturalness of the speech is achieved, (2) the converted speech indeed sounds like the target speaker, and not like the original speaker (similarity of speaker characteristics).

Naturalness is usually evaluated using subjective tests, though objective metrics such as the Mel-cepstral distortion (MCD) [40], Modulation Spectrum (MS) [41], Global Variance (GV) [42] or the Word Error Rate (WER) are often used during prototyping . The most common subjective way to evaluate the naturalness of a speech sample is the Mean Opinion Score (MOS) evaluation. The MOS is a 5-point Likert scale, which was originally introduced for measuring the quality of telephone speech [43], but later adopted as a standard for voice conversion naturalness evaluation [44].

Even though the synthesised speech sounds natural, it could happen that the voice characteristics had not moved away from the source speaker. To avoid this issue, the speaker similarity is tested using an AB-test setup. During the experiment, listeners are presented with synthesised and real utterances in random pairwise combinations. The listeners have to indicate on a 4-point Likert scale, whether they think the speech samples are from the same or a different speaker.

Naturalness testing is often criticised in the speech synthesis literature [45, 46, 47, 48], however, little to no alternatives have been provided [46]. The main source of criticism is that several factors irrelevant to the synthesis quality, such as gender, age, or attitude towards speech synthesis technology, seem to influence perceived naturalness [49]. An additional criticism is that current naturalness tests provide a single global measure of the naturalness of the synthesised speech [47]. This means that MOS is largely non-informative regarding what component of the speech should be improved.

The situation is even harder in the case of pathological speech synthesis because even natural pathological speech negatively affects MOS scores, as we will see in Chapters 5 and 6. In a development scenario, a pathological VC system that receives better MOS scores than the reference pathological speech might be just synthesising healthy speech. Reversely, a VC system mimicking the pathology, although in an exaggerated manner, would likely produce a MOS score that is a lot lower than that of the reference pathological speech.

Making sure that the synthesised sample sounds like the pathological speaker also poses some challenges in the case of pathological speech. Certain pathological speech conditions are associated with the loss of identifiable vocal characteristics, (e.g., laryngectomee speech) which makes speaker identity evaluation ill-posed in the case of pathological speakers. We will see an example of this in Chapter 5, where we show that this can be indeed an issue in practice.

1.5. CONTRIBUTION OF THE PRESENT THESIS

In this thesis, we will contribute to three different speech technology tasks on pathological and atypical speech: automatic speech recognition for atypical speakers (*bias*); automatic *ecologically valid* objective speech severity evaluation for oral cancer speakers; and voice conversion methods for pathological speech that alleviates the *disentanglement* and *evaluation* issues mentioned above.

The main research questions (RQ) investigated in this thesis are the following:

RQ1 To what extent is ASR performance diminished in the case of atypical and pathological speakers? What are the main reasons for the worse recognition performance and how can we address/alleviate these? Specifically, we will investigate whether WER is influenced by pronunciation/articulation, noise and severity of speech.

Chapter 2 shows that state-of-the-art Dutch speech recognisers are already *biased* against speaker whom we do not consider pathological, i.e., children, old people, and non-native speakers, but who have speaker characteristics that are deviant from the typical population of speakers on which ASR systems are trained, i.e., native adult speakers without a strong accent or speech pathology. We quantify the *bias* in performance against these speaker groups compared to the typical speakers in the training material and look into the possible source of the recognition errors on the phoneme level. **Chapter 3** develops and compares systems for the recognition of oral cancer speech in spontaneous scenarios, and performs a similar articulatory analysis as **Chapter 2**, albeit with greater detail.

RQ2 How well can we automatically predict the severity of spontaneous oral cancer speech? Can we explain the decisions made by these systems?

Chapter 4 uses the ASR systems developed in **Chapter 3** and compares them with acoustic feature-, ASR-, and comparison-based systems to assess whether it is possible to predict pathology severity ratings automatically from speech samples. We tested and developed several systems that have an explainable component within them on the task of severity prediction in an *ecologically valid* scenario.

RQ3 Is it possible to convert healthy speech to pathological speech through voice conversion while preserving the identity of the healthy speaker, and achieving naturalness comparable to real dysarthric speech?

Chapter 5 proposes a VC setup using a pathological source utterance, which is customised to another pathological speaker's voice characteristics. Using a pathological source utterance ensures that the time-variant characteristics of the speech are already changed correctly, therefore only the time-invariant characteristics of the speech, i.e., the speaker identity, have to be changed (*disentanglement*). This also alleviates the issue in the naturalness *evaluation*, as the model is not directly trained to increase severity of the speech. The chapter also provides the first evidence that synthesis quality (naturalness) is influenced by the severity of the speech.

Chapter 6 builds on the framework outlined in **Chapter 5** to alleviate the need for pathological input speech, creating pathological speech from healthy speech, instead of pathological speech. The *evaluation* issues are further alleviated by using a two-pronged evaluation on both stages of the model, with a mixture of objective and subjective methods.

Finally, in the **Discussion**, we are going to revisit and answer the research questions. Moreover, we are going to take a step back, and see what these results mean for the near future of accessible speech technology.

1.6. AUTHOR CONTRIBUTIONS

The following papers are completed by the author of this thesis while pursuing the Ph.D. at the University of Amsterdam.

- Feng, S., **Halpern, B. M.**, Kudina, O. & Scharenborg, O. (2022). Towards inclusive automatic speech recognition. Submitted to Computer Speech and Language. **[Chapter 2]** The PhD candidate contributed to the writing, and evaluation of the experiments.

- **Halpern, B. M.***, Feng, S.*, van Son, R., van den Brekel, M., & Scharenborg, O. (2022). Low-resource automatic speech recognition and error analyses of oral cancer speech. Speech Communication. **[Chapter 3]** The PhD candidate contributed to the data collection, experimental design, writing, and evaluation of the experiments.

- **Halpern, B. M.**, Feng, S., van Son, R., van den Brekel, M., & Scharenborg, O. (2022). Automatic evaluation of spontaneous oral cancer speech using ratings from naive listeners. Submitted to Speech Communication. **[Chapter 4]** The PhD candidate contributed to the data collection, implementation, experimental design, writing and evaluation of the experiments.

- **Halpern, B. M.***, Illa M.*, van Son, R., Moro-Velázquez, L., & Scharenborg, O. (2021). Pathological voice adaptation with autoencoder-based voice conversion. In 11th ISCA Speech Synthesis Workshop (pp. 19-24). ISCA. **[Chapter 5]** The PhD candidate contributed to the supervision, experimental design, and evaluation of the project. Marc Illa was a master student supervised by the PhD candidate.

- **Halpern, B. M.***, Huang, W.C.*, Violeta, L. P., Scharenborg, O., & Toda, T. (2022, May). Towards identity preserving normal to dysarthric voice conversion. In ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP) (pp. 6672-6676). IEEE. **[Chapter 6]** The PhD candidate contributed to the writing, the idea and the evaluation of the project.

We note that **Chapter 2** is based on a publication where the PhD candidate is not the first author. However, this publication is also included in this thesis due to substantial contribution to the evaluation of the experiments, and the writing of the paper. **Chapters 3, 5** and **6** have shared first authorship.

BIBLIOGRAPHY

[1] O. Russakovsky, J. Deng, H. Su, J. Krause, S. Satheesh, S. Ma, Z. Huang, A. Karpathy, A. Khosla, M. Bernstein *et al.*, "Imagenet large scale visual recognition challenge," *International Journal of Computer Vision*, vol. 115, no. 3, pp. 211–252, 2015.

[2] T. Parcollet and M. Ravanelli, "The energy and carbon footprint of training end-to-end speech recognizers," *HAL Archives Ouvertes Preprint Hal-03190119f*, 2021.

[3] R. Choenni, E. Shutova, and R. van Rooij, "Stepmothers are mean and academics are pretentious: What do pretrained language models learn about you?" *arXiv preprint arXiv:2109.10052*, 2021.

[4] J. Buolamwini and T. Gebru, "Gender shades: Intersectional accuracy disparities in commercial gender classification," in *Conference on Fairness, Accountability and Transparency*. PMLR, 2018, pp. 77–91.

[5] E. M. Bender, T. Gebru, A. McMillan-Major, and S. Shmitchell, "On the dangers of stochastic parrots: Can language models be too big?" in *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, 2021, pp. 610–623.

[6] A. Hinsvark, N. Delworth, M. Del Rio, Q. McNamara, J. Dong, R. Westerman, M. Huang, J. Palakapilly, J. Drexler, I. Pirkin *et al.*, "Accented speech recognition: A survey," *arXiv preprint arXiv:2104.10747*, 2021.

[7] K. Rosen and S. Yampolsky, "Automatic speech recognition and a review of its functioning with dysarthric speech," *Augmentative and Alternative Communication*, vol. 16, no. 1, pp. 48–60, 2000.

[8] T. Pommée, M. Balaguer, J. Mauclair, J. Pinquier, and V. Woisard, "Intelligibility and comprehensibility: A delphi consensus study," *International Journal of Language & Communication Disorders*, vol. 57, no. 1, pp. 21–41, 2022.

[9] N. Miller, "Measuring up to speech intelligibility," *International Journal of Language & Communication Disorders*, vol. 48, no. 6, pp. 601–612, 2013.

[10] AHSA, "Policy." [Online]. Available: https://www.asha.org/policy/sp2016-00343/

[11] J. R. Duffy, *Motor Speech Disorders: Substrates, Differential Diagnosis and Management*, 2005.

[12] Y. Maryn and K. Debo, "Is perceived dysphonia related to perceived healthiness?" *Logopedics Phoniatrics Vocology*, vol. 40, no. 3, pp. 122–128, 2015.

[13] G. Henningsson, D. P. Kuehn, D. Sell, T. Sweeney, J. E. Trost-Cardamone, and T. L. Whitehill, "Universal parameters for reporting speech outcomes in individuals with cleft palate," *The Cleft Palate-Craniofacial Journal*, vol. 45, no. 1, pp. 1–17, 2008.

[14] H. Kim, M. Hasegawa-Johnson, A. Perlman, J. Gunderson, T. S. Huang, K. Watkin, and S. Frame, "Dysarthric speech database for universal access research," in *Ninth Annual Conference of the International Speech Communication Association*, 2008.

[15] C. Bhat, B. Vachhani, and S. K. Kopparapu, "Automatic assessment of dysarthria severity level using audio descriptors," in *2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2017, pp. 5070–5074.

[16] A. A. Joshy and R. Rajan, "Automated dysarthria severity classification using deep learning frameworks," in *2020 28th European Signal Processing Conference (EUSIPCO)*. IEEE, 2021, pp. 116–120.

[17] T. Lee, Y. Liu, Y. T. Yeung, T. K. Law, and K. Y. Lee, "Predicting Severity of Voice Disorder from DNN-HMM Acoustic Posteriors," in *Interspeech*, 2016, pp. 97–101.

[18] F. L. Darley, A. E. Aronson, and J. R. Brown, "Differential diagnostic patterns of dysarthria," *Journal of Speech and Hearing Research*, vol. 12, no. 2, pp. 246–269, 1969.

[19] R. Savica, B. R. Grossardt, W. A. Rocca, and J. H. Bower, "Parkinson disease with and without dementia: a prevalence study and future projections," *Movement Disorders*, vol. 33, no. 4, pp. 537–543, 2018.

[20] G. Moya-Galé and E. S. Levy, "Parkinson's disease-associated dysarthria: prevalence, impact and management strategies," 2019.

[21] R. DeRosier and R. S. Farber, "Speech recognition software as an assistive device: a pilot study of user satisfaction and psychosocial impact," *Work*, vol. 25, no. 2, pp. 125–134, 2005.

[22] L. Moro-Velazquez, J. Cho, S. Watanabe, M. A. Hasegawa-Johnson, O. Scharenborg, H. Kim, and N. Dehak, "Study of the Performance of Automatic Speech Recognition Systems in Speakers with Parkinson's Disease," in *Proc. Interspeech 2019*, 2019, pp. 3875–3879. [Online]. Available: http://dx.doi.org/10.21437/Interspeech.2019-2993

[23] V. Young and A. Mihailidis, "Difficulties in automatic speech recognition of dysarthric speakers and implications for speech-based applications used by the elderly: A literature review," *Assistive Technology : The Official Journal of RESNA*, vol. 22, pp. 99–112; quiz 113, 06 2010.

[24] C. Llewellyn, M. McGurk, and J. Weinman, "How satisfied are head and neck cancer (hnc) patients with the information they receive pre-treatment? results from the satisfaction with cancer information profile (scip)," *Oral Oncology*, vol. 42, no. 7, pp. 726–734, 2006. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S1368837505003313

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

[26] R. I. Zraick, G. B. Kempster, N. P. Connor, S. Thibeault, B. K. Klaben, Z. Bursac, C. R. Thrush, and L. E. Glaze, "Establishing validity of the consensus auditory-perceptual evaluation of voice (CAPE-V)," *American Journal of Speech-Language Pathology*, vol. 20, no. 1, p. 14, 2011.

[27] J. Oates, "Auditory-perceptual evaluation of disordered voice quality," *Folia Phoniatrica et Logopaedica*, vol. 61, no. 1, pp. 49–56, 2009.

[28] M. Schuster, A. Maier, T. Haderlein, E. Nkenke, U. Wohlleben, F. Rosanowski, U. Eysholdt, and E. Nöth, "Evaluation of speech intelligibility for children with cleft lip and palate by means of automatic speech recognition," *International Journal of Pediatric Otorhinolaryngology*, vol. 70, no. 10, pp. 1741–1747, 2006.

[29] M. Windrich, A. Maier, R. Kohler, E. Nöth, E. Nkenke, U. Eysholdt, and M. Schuster, "Automatic quantification of speech intelligibility of adults with oral squamous cell carcinoma," *Folia Phoniatrica et Logopaedica*, vol. 60, no. 3, pp. 151–156, 2008.

[30] M. Schuster, T. Haderlein, E. Nöth, J. Lohscheller, U. Eysholdt, and F. Rosanowski, "Intelligibility of laryngectomees' substitute speech: automatic speech recognition and subjective rating," *European Archives of Oto-Rhino-Laryngology and Head & Neck*, vol. 263, no. 2, pp. 188–193, 2006.

[31] T. Bressmann, R. Sader, T. L. Whitehill, and N. Samman, "Consonant intelligibility and tongue motility in patients with partial glossectomy," *Journal of Oral and Maxillofacial Surgery*, vol. 62, no. 3, pp. 298–303, 2004.

[32] T. Bressmann, H. Jacobs, J. Quintero, and J. C. Irish, "Speech outcomes for partial glossectomy surgery: Measures of speech articulation and listener perception," *Head and Neck Cancer*, vol. 33, no. 4, p. 204, 2009.

[33] A. Maier, T. Haderlein, U. Eysholdt, F. Rosanowski, A. Batliner, M. Schuster, and E. Nöth, "Peaks–a system for the automatic evaluation of voice and speech disorders," *Speech Communication*, vol. 51, no. 5, pp. 425–437, 2009.

[34] T. Tienkamp, R. van Son, and B. M. Halpern, "Objective speech outcomes after surgical treatment for oral cancer: An acoustic analysis of a spontaneous speech corpus containing 32.850 tokens," *Submitted to Journal of Communication Disorders*, 2022.

[35] G. P. Prajapati, D. K. Singh, P. P. Amin, and H. A. Patil, "Voice privacy using cyclegan and time-scale modification," *Computer Speech & Language*, vol. 74, p. 101353, 2022.

[36] Z. Qian, H. Niu, L. Wang, K. Kobayashi, S. Zhang, and T. Toda, "Mandarin electrolaryngeal speech enhancement based on statistical voice conversion and manual tone control," in *2021 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC)*. IEEE, 2021, pp. 546–552.

[37] W.-C. Huang, K. Kobayashi, Y.-H. Peng, C.-F. Liu, Y. Tsao, H.-M. Wang, and T. Toda, "A preliminary study of a two-stage paradigm for preserving speaker identity in dysarthric voice conversion," *arXiv preprint arXiv:2106.01415*, 2021.

[38] J. Wakefield, "Deepfake presidents used in russia-ukraine war." [Online]. Available: https://www.bbc.com/news/technology-60780142

[39] T. Kinnunen, M. Sahidullah, H. Delgado, M. Todisco, N. Evans, J. Yamagishi, and K. A. Lee, "The asvspoof 2017 challenge: Assessing the limits of replay spoofing attack detection," 2017.

[40] R. Kubichek, "Mel-cepstral distance measure for objective speech quality assessment," in *Proceedings of IEEE Pacific Rim Conference on Communications Computers and Signal Processing*, vol. 1. IEEE, 1993, pp. 125–128.

[41] S. Takamichi, T. Toda, A. W. Black, and S. Nakamura, "Modulation spectrum-based post-filter for gmm-based voice conversion," in *Signal and Information Processing Association Annual Summit and Conference (APSIPA), 2014 Asia-Pacific*. IEEE, 2014, pp. 1–4.

[42] T. Toda, A. W. Black, and K. Tokuda, "Voice conversion based on maximum-likelihood estimation of spectral parameter trajectory," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 15, no. 8, pp. 2222–2235, 2007.

[43] I. T. Union, "Methods for subjective determination of transmission quality," 1996.

[44] B. Sisman, J. Yamagishi, S. King, and H. Li, "An overview of voice conversion and its challenges: From statistical modeling to deep learning," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 29, pp. 132–157, 2020.

[45] P. Wagner, J. Beskow, S. Betz, J. Edlund, J. Gustafson, G. Eje Henter, S. Le Maguer, Z. Malisz, Éva Székely, C. Tånnander, and J. Voße, "Speech Synthesis Evaluation — State-of-the-Art Assessment and Suggestion for a Novel Research Program," in *Proc. 10th ISCA Workshop on Speech Synthesis (SSW 10)*, 2019, pp. 105–110.

[46] R. C. Streijl, S. Winkler, and D. S. Hands, "Mean opinion score (mos) revisited: methods and applications, limitations and alternatives," *Multimedia Systems*, vol. 22, no. 2, pp. 213–227, 2016.

[47] E. Gutierrez, P. Oplustil-Gallegos, and C. Lai, "Location, Location: Enhancing the Evaluation of Text-to-Speech synthesis using the Rapid Prosody Transcription Paradigm," in *Proc. 11th ISCA Speech Synthesis Workshop (SSW 11)*, 2021, pp. 25–30.

[48] S. King, "Measuring a decade of progress in text-to-speech," *Loquens*, vol. 1, no. 1, pp. e006–e006, 2014.

[49] S. Jumisko-Pyykkö and J. Häkkinen, "Profiles of the evaluators: impact of psychographic variables on the consumer-oriented quality assessment of mobile television," in *Multimedia on Mobile Devices 2008*, vol. 6821. International Society for Optics and Photonics, 2008, p. 68210L.

2

TOWARDS INCLUSIVE AUTOMATIC SPEECH RECOGNITION

Practice and recent evidence show that state-of-the-art (SotA) automatic speech recognition (ASR) systems do not perform equally well for all speaker groups. Many factors can cause this bias against different speaker groups. This paper, for the first time, systematically quantifies and finds speech recognition bias against gender, age, regional accents and non-native accents, and investigates the origin of this bias by investigating bias cross-lingually (i.e., Dutch and Mandarin) and for two different SotA ASR architectures (a hybrid DNN-HMM and an attention based end-to-end (E2E) model) through a phoneme error analysis. The results show that only a fraction of the bias can be explained by pronunciation differences between speaker groups, and that in order to mitigate bias, language- and architecture specific solutions need to be found.

2.1. INTRODUCTION

Automatic speech recognition (ASR) is increasingly used, in, e.g., emergency response centres, domestic voice assistants, and search engines. Because of the paramount relevance spoken language plays in our lives, it is critical that ASR systems are able to deal with the variability in the way people speak (e.g., due to speaker differences, demographics, and differently abled speakers).

State-of-the-art (SotA) ASR systems are based on deep neural networks (DNNs). DNNs are often considered to be a harbour of objectivity because they follow a clear path against the set parameters applied to the provided dataset. Although studies on bias in ASR are only nascent, practice and recent evidence are already troubling, suggesting that the SotA ASR systems do not recognise the speech of everyone equally well. This evidence ranges from anecdotal (e.g., the smart speaker of author O.S. does not

This chapter has been submitted as: Feng, S., Halpern, B. M., Kudina, O. & Scharenborg, O. (2022). Towards inclusive automatic speech recognition. Computer Speech and Language. The PhD candidate contributed to the writing, and evaluation of the experiments.

recognise the speech of her 9-year-old daughter) to research- and policy-oriented. For instance, ASR systems have been shown to struggle with speech variance due to gender, age, speech impairment, race, and accents. Studies across languages have repeatedly found recognition bias between genders, predominantly favouring female speakers (Arabic [1], English [2, 3, 4], and French [3]), while male speech was best recognised in other studies (French [5], English [6]), although a follow-up study to the latter study found no difference between genders [7] nor was a difference found in [5]. It should be noted that these studies do not include transgender and non-binary speakers.

Speakers younger than 30 years of age are better recognised than those older than 30 years [1], while the recognition of child speech is more challenging than that for adult speech, due to children's shorter vocal tracts, slower and more variable speaking rate and inaccurate articulation [8]. A speech impairment, e.g., due to dysarthria [9], stroke survival, oral cancer [10] or cleft lip and palate [11], is known to cause many problems for standard ASR systems. Recent studies further demonstrate how voice assistants perpetuate a racial divide by misrecognising the speech of African American speakers more often than of white speakers [2, 7]. Finally, ASR systems are typically trained on speech from native speakers of a "standard" variant of that language, inadvertently discriminating not only the speech of non-native speakers [12, 13] but also that of speakers of regional or sociolinguistic variants of the language (English [2, 6, 7], Arabic [1]).

There are many factors that can cause this bias, and different locations in the ASR system where these factors manifest themselves. Such bias-inducing factors, for instance, include, 1) under-representation of the speaker group in the training data (i.e., the composition of the training data). This leads to acoustic models (AMs) that will not be able to capture the pronunciations of that speaker group well. 2) Within-group variability: Even if the ASR is trained only on speech of the underrepresented group, recognition performance is often found to be worse due to the large variability both in the pronunciation and in language use within the speaker group (e.g., [2, 7, 8]). 3) The transcriptions can be biased. Anecdotal evidence (from author B.M.H. on the Jasmin-CGN corpus [14] suggests that production errors of children are corrected ("normalised" towards what should have been said) in a more lenient way than those of non-native adult speakers (transcriptions tend to be more verbatim, including restarts). Moreover, transcriptions might be less accurate because the annotators have less experience with the type of speech. 4) Across-group variability: A speaker group that has a dialect that deviates significantly from that of the other speaker groups in the training data is usually recognised worse [15, 16]. 5) Not all speaker groups might have access to equally high-quality recording equipment. 6) Possibly, bias can be due to the specific ASR architectures, of which there are two main categories in current ASR: end-to-end (E2E) and hybrid Deep Neural Network (DNN)-hidden Markov model (DNN-HMM), and algorithms used in ASR system development. 7) Bias also creeps in far before the datasets are collected and deployed, e.g., when framing the problem, preparing the data and collecting it (e.g., [17]). Most of these factors will have their impact on the acoustic model (AM), e.g., leading to a mismatch with the trained AM. However, deviant language use will also have an effect on the language model.

Our goal is to create inclusive ASR, i.e., ASR for everyone, irrespective of how one speaks or the language one speaks. As first and crucial steps towards this larger goal,

Table 2.1: Hours of speech data and numbers of speakers in the CGN training and test sets for Dutch.

	Training		BN test		CTS test	
	#Hrs	#Spks	#Hrs	#Spks	#Hrs	#Spks
All	423	2863	0.4	4	1.8	25
Female	193	1185	0.2	1	0.8	12
Male	230	1678	0.2	3	1.0	13

here, we systematically quantify bias against different speaker groups, and investigate the origin of this bias by investigating bias 1) in two different speaking styles in order to answer the question whether the size of the bias is influenced by the speaking style of the person, 2) cross-lingually in two vastly different languages (non-tonal Dutch and tonal Mandarin) in order to answer the question whether bias is language dependent, and 3) for different SotA ASR architectures (a hybrid DNN-HMM and an attention based E2E model) in order to answer the question whether bias is dependent on the ASR architecture. The results will allow us to work towards proactive bias-mitigation in ASR systems.

Prior work in the literature typically focused on one to three speaker groups or dimensions, here we will investigate possible bias against gender, age, regional accents, and non-native accents. In our search for the origin of the bias, we carry out an analysis of which sounds are particularly prone to misrecognition. Error patterns might be indicative of particular problems that lead to bias, e.g. misrecognised vowels or voiced final obstruents misrecognised as their unvoiced counterparts might be indicative of regional or non-native speech.

2.2. SPEECH DATABASE SELECTION AND DESIGN

In order to be able to quantify bias for different speaker groups, we are crucially dependent on the meta-data available in the speech databases. We therefore carefully selected and curated our speech databases. For Mandarin, we only found databases that allowed us to investigate bias against gender and regional accents.

2.2.1. DUTCH CORPORA

DUTCH SPOKEN CORPUS (CGN)

The CGN corpus [18] is used to train the Dutch SotA ASR systems. CGN contains Dutch recordings spoken by 1185 female and 1678 male speakers (age range 18-65 years old) from all over the Netherlands (NL) and Flanders (FL, in Belgium (BE)). It contains 14 different speaking styles. In this study, only CGN data from NL is used for training, while both Dutch from NL and FL are used for testing. We used the standard training and test sets [19]. Two test sets were used, one for each speaking style: broadcast news (BN) and conversational telephone speech (CTS). All recordings were first pre-processed by cutting the speech signals into smaller chunks and removing the silence chunks. Table 2.1 presents detailed information about the CGN training and test sets after the pre-processing steps.

Table 2.2: Number of native female and male speakers of Dutch per age group per NL region (D-) and for FL (F-) in the Jasmin-CGN corpus.

Region	W	T	N	S	FL
DC/FC	0, 0	15, 14	11, 11	9, 11	23, 19
DT/FT	9, 11	2, 2	10, 10	10, 9	22, 21
DOA/FOA	13, 5	9, 8	13, 4	10, 6	21, 16

JASMIN-CGN CORPUS

The Jasmin-CGN corpus [14], which is an extension of the CGN corpus, is used to evaluate the Dutch ASR systems' bias against gender, age, regional and non-native accent[1]. We use the speech from the following groups: (1) DC: Dutch children; age 7–11 years; (2) DT: Dutch teenagers; age 12–16; (3) DOA: Dutch older adults; age 65+; (4) NNC: non-native Dutch speaking children; age 7–16 years (28 female and 25 male speakers); (5) NNA: non-native adults; age 18–60 (28 female and 17 males speakers); with a wide range of native languages. The adults have different levels of proficiency of Dutch according to the Common European Framework (CEF; A1 the lowest): A1 (4 females, 6 males), A2 (18 females, 7 males), B1 (6 females, 3 males), B2 (1 male). The speakers come from four different regions in NL: W: West, T: Transitional, N: North, S: South. Moreover, we tested the ASR trained on NL Dutch on the speech of (1) FC: Flemish children; age 7–11; (2) FT: Flemish teenagers; age 12–16; (3) FOA: Flemish older adults; age 65+.

Table 2.2 shows the number of speakers broken down by gender (female, male[2]) for each age group and each (NL or FL) region, excluding the non-native speakers for which this information was not available. The Jasmin-CGN corpus consists of read speech and human-machine interaction (HMI) speech, both of which are used in the experiments. The number of hours for each region and age group ranges from 0.2h (DT from region T) to 2.0h (female FT) and from 0.2h (several speaker groups from region S) to 0.9h (FOA) for HMI speech. The number of hours of speech of the non-native speakers ranged from 0.2h (B2 male speaker) to 0.8h (A1 and A2 male speakers for read speech and from 0.1h (A1 female speakers) to 1.1h (A2 female speakers) for HMI speech.

2.2.2. MANDARIN CORPUS

The MagicData Read Speech Corpus [20] is an open-source Mandarin speech corpus consisting of 755 hours of Mandarin recordings spoken by adult speakers (age range 18-55 years old) from seven regions from all over mainland China: Northern Guan (NG), Southern Guan (SG), Gan (GA), Min (MI), Wu (WU), Xiang (XI) and Yue (YU)[3]. In the seven regions, speakers from NG and SG use a variety of Mandarin as their local languages, while local languages in GA, MI, WU, XI and YU are non-Mandarin Sinitic lan-

[1] The recording conditions of CGN and Jasmin-CGN are different, which might lead to an ASR performance deterioration on Jasmin-CGN compared to CGN. However, in our bias investigations we only compare speaker groups within the Jasmin-CGN dataset, avoiding this potential problem.

[2] Please note that the meta-data only provides information regarding these two genders.

[3] Hakka is also a major Sinitic language; it is mainly spoken in the provinces of Guangdong, Guangxi and Fujian. However, Hakka is not included in this paper, as region information for speakers in MagicData is only available at the province level, making Hakka indistinguishable from Yue (Guangdong, Guangxi) and Min (Fujian).

Table 2.3: Hours of speech data and number of (male/female) speakers in the MagicData training and test sets. The test data is also broken down by accent region.

Set		All		Female		Male	
		#Hrs	#Spks	#Hrs	#Spks	#Hrs	#Spks
Training		680.1	968	366.9	516	313.2	452
Test	All	52.1	78	26.5	37	25.6	41
	NG	11.1	16	8.4	11	2.7	5
	SG	8.7	12	2.4	3	6.3	9
	GA	6.5	10	2.8	4	3.7	6
	MI	7.0	10	3.8	5	3.2	5
	WU	5.6	10	2.3	4	3.2	6
	XI	6.5	10	2.9	4	3.6	6
	YU	6.7	10	4.0	6	2.8	4

guages. The supplementary information 2.6 provides the mapping from a Chinese province to its accent region.

We followed the standard training, development and test data partitioning in MagicData, but with two necessary modifications: (1) The original test set does not contain 10 speakers for all accent regions. In order to avoid the results being dependent on the characteristics of individual speakers, we empirically set the minimum number of test speakers in every Chinese accent region at 10. To that end, speakers from the original training set were randomly selected and moved to the test set. (2) The original test set did not contain any female speech for NG and SG. In order to balance gender, the female speakers from the NG and SG regions in the original development set were moved to the test set. There is no speaker overlap between the training and test sets. Table 2.3 shows the number of hours of speech and test speakers broken down by gender and by Chinese accent regions in our test data.

2.2.3. EXPERIMENTS AND EVALUATION

In our experiments on Dutch, the potential bias due to gender, age, regional and non-native accents is quantified for read speech and HMI speech separately. For Mandarin, the bias against gender and regional accents is quantified for read speech only.

We quantify the bias of ASR systems for Dutch on the Jasmin-CGN corpus and for Mandarin on the MagicData corpus. We define bias as the difference in WER between the different speaker groups within each of the investigated dimensions, and it is computed by subtracting the lowest (=best) WER from the WER of each speaker group in the dimension. Moreover, for Dutch, for all dimensions, we split by age group. One-way analyses of variance (ANOVA) were carried out comparing the WER of each speaker group within each of the dimensions to investigate the significance of the bias, with per-speaker WER as the dependent variable, and each of the dimensions as the independent variable.

In order to understand the source of the bias, we computed phoneme error rates (PER) for each individual phoneme to investigate whether certain phonemes are prone to misrecognitions. The PER is calculated similarly to the calculation of the WER but using phoneme-level transcriptions (converted from word-level transcriptions using lexi-

cons) of the reference and hypothesised word sequences[4].

2.2.4. THE STATE-OF-THE-ART HYBRID AND E2E ASR SYSTEMS

For the experiments, we used a SotA factorised TDNN [21] (TDNNF) implemented using Kaldi [22][5] as our hybrid model and a conformer-based encoder-decoder model implemented using ESPnet as our E2E model, one for Dutch and one for Mandarin [24]. For both architectures, the same training material and MFCC acoustic feature representations were used. The language model (LM) in the hybrid ASR system is an RNNLM [25]. The same RNNLM was used during E2E ASR decoding as shallow fusion [26] to the conformer ASR model.[6].

Since there are no standard read speech and HMI test sets in CGN, the TDNNF and E2E were first evaluated on the CGN standard BN and CTS test sets for reference. On BN speech, the TDNNF system (6.3%) slightly outperformed the E2E system (6.6%), while the E2E system outperformed the TDNNF (21.6% vs. 23.9%) on CTS speech. Details of the in-domain WER results are listed in Table 2.8.

2.3. QUANTIFYING BIAS

2.3.1. BIAS IN STATE-OF-THE-ART ASRS FOR DUTCH

BIAS AGAINST GENDER

Overall, female speech was recognised similarly or better than male speech (see Table 2.10 in the supporting materials for the WER breakdown for gender, age and non-nativeness). Table 2.4a lists the size of the bias against male speakers compared to female speakers split for native and non-native speakers and for the different age groups, for the hybrid and E2E systems and the read speech and HMI speech test sets.

The native speaker groups – for the hybrid models, we only observe bias for two cases: for read speech, male teenagers are significantly worse recognised than female teenagers ($F(1,61)=5.543$, $p=.022$), while for HMI speech, female children are significantly worse recognised than male children ($F(1,69)=4.316$, $p=.041$). The E2E ASR is more prone to bias. We observed a statistically significant bias against male speakers for teenagers (read speech: $F(1,61)=7.953$, $p=.006$; HMI speech: ($F(1,61)=4.036$, $p=.049$)) and older adults (read speech: $F(1,66)=4.122$, $p=.046$; HMI speech: ($F(1,66)=6.350$, $p=.014$)) in both speech styles. No biases were observed for the **non-native speaker groups**.

Both architectures thus exhibited a bias against male speakers, however this bias was much less for the hybrid model compared to the E2E model. No gender bias was observed for the non-native listeners. This finding could however be due to the relatively

[4] Source code of the analysis method can be found at: https://github.com/karkirowle/relative_phoneme_analysis.

[5] A preliminary experiment compared a TDNN-BLSTM [23] and a factorised TDNN [21] (TDNNF) model, both were implemented using Kaldi [22], used the same training material and the same MFCC acoustic features. Although the TDNN-BLSTM outperformed the TDNNF system on the in-domain CGN BN set, the TDNN-BLSTM performed worse than the TDNNF on the out-domain Jasmin-CGN corpus, our test corpus. For Mandarin, we also observed that the TDNNF model outperformed the TDNN-BLSTM model. Therefore, in our experiments, we used the TDNNF system. Details of the in-domain and out-domain WER results for Dutch are listed in Tables 2.8 and 2.9.

[6] Open-source code to replicate our experiments: https://github.com/syfengcuhk/jasmin.

Table 2.4: Bias sizes for the TDNNF hybrid and conformer E2E ASR systems for gender (2.4a), age (2.4b), and regional accents (2.4c) split by age group on the Jasmin-CGN read and HMI speech. * = p < .05; ** = p < .01; *** = p < .001.

(a) Bias against gender. Female speech was recognised best.

	Read		HMI	
	Hybrid	E2E	Hybrid	E2E
DC	0.3	-0.5	-3.0*	-0.4
DT	2.5*	3.2**	1.5	3.7*
DOA	5.1	5.3*	5.1	6.5*
NNC	1.1	1.8	1.0	4.4
NNA	0.4	1.3	3.7	5.6

(b) Bias against age group. For the Dutch speakers, teenagers were recognised best; for the non-native speakers, child speech was recognised best.

	Read		HMI	
	Hybrid	E2E	Hybrid	E2E
DC	11.8***	12.2***	7.2***	7.2***
DOA	4.6**	3.9*	7.8***	8.2***
NNA	1.6	1.9	1.2	1.6

(c) Bias against native regional accents. Bias size numbers are calculated by subtracting the lowest WER (first field within brackets) from the highest WER (second field within brackets) among five regions (including FL) within an age group. 'Avg' indicates the average bias size over all age groups.

	Read		HMI	
	Hybrid	E2E	Hybrid	E2E
Regions: W, T, N, S, FL				
DC/FC	11.5*** (T,FL)	12.6*** (T,FL)	20.5*** (N,FL)	19.6*** (N,FL)
DT/FT	16.4*** (N,FL)	20.9*** (N,FL)	15.8* (T,FL)	19.9** (T,FL)
D/FOA	11.7** (N,S)	12.2** (N,S)	13.1* (N,S)	13.8*** (N,FL)
Avg.	13.2	15.2	16.5	17.8

high WERs for the non-native speaker groups. These results add to a growing set of findings that male and female speech are not recognised equally well [2, 6, 1, 3, 4].

BIAS AGAINST AGE

Overall, for the native speakers, speech from teenagers was recognised best, followed by that of older adults, while child speech was recognised worst (see Table 2.10 for the WER breakdown for gender, age and non-nativeness). For the non-native speakers, child speech was recognised better than that of adult speakers. Table 2.4b lists the size of the bias against native children's and older adults' speech (top rows) and against non-native adults' speech (bottom row), for read and HMI speech, and the hybrid and E2E models, separately.

The native speakers groups – we observe substantial age bias: speech from teenagers was found to be significantly better recognised than that of children for both models and both speaking styles (hybrid ASR on read speech: $(F(1,132)=87.158, p<.001)$; on HMI speech: $(F(1,132)=19.425, p<.001)$; E2E ASR on read speech: $(F(1,132)=88.815, p<.001)$; on HMI speech: $(F(1,132)=25.691, p<.001)$) and significantly better recognised than speech from older adults for both models and both speaking styles (hybrid ASR on read speech:

(F(1,129)=7.573, p=.007); on HMI speech: (F(1,129)=15.804, p<.001); E2E ASR on read speech: (F(1,129)=6.533, p=.012); on HMI speech: (F(1,129)=18.935, p<.001)).

The age bias size, though, is different for the two speech styles: the bias against children's speech is smaller for HMI than for read speech, while the bias against older adults' speech is larger for HMI than for read speech. Informal listening to a few of the child speakers' recordings suggests that the smaller bias for HMI speech is due to a higher WER on read speech, and could be due to high volume and disfluency/hesitations in the children's read speech. No bias was observed for the **non-native adult speakers** compared to the speech of that of non-native children.

In conclusion, both architectures exhibited a (large) age bias against children's and older adults' speech for native speakers of Dutch, while no age bias was observed for the non-native speakers. The size of the bias seems to be similar for the two architectures. The problems of the ASR with recognising children's speech can be explained by the large difference in children's speech and adults' speech [8] which leads to a large mismatch of the children's speech with the AM. The worse recognition of the older adults' speech, especially those over 75 y/o, is likely due to a less well articulation.

BIAS AGAINST NON-NATIVE ACCENTS

The speech of native speakers was recognised better than that of non-native speakers of Dutch (Table 2.10). The bias against non-native accents is significant for both speaking styles and both architectures (hybrid ASR on read speech: Size = 23.1; F(1,298)=282.851,p<.001; on HMI speech: Size = 13.9; F(1,298)=126.716,p<.001; E2E ASR on read speech: Size = 24.5; F(1,298)=344.457,p<.001; on HMI speech: Size = 16.7; F(1,298)=197.807,p<.001). The hybrid system seemed to exhibit a smaller bias against the non-native speakers than the E2E architecture.

These results are in line with the qualitative findings reported in [12, 13]. Non-native speakers typically have an accent, meaning that the match with the AM is worse than that of native speakers. For the non-native speakers, on average, the WER results by both models showed a decrease when CEF level increases (see Table 2.11; except for the one B1 speaker for read speech). This is in line with the intuition that non-native speakers with a higher CEF level tend to speak Dutch better than those with a lower level.

BIAS AGAINST REGIONAL ACCENTS

Overall, there is a large variety in the recognition performance of the speech from the different accent regions in the Netherlands and Flanders, with speech from Flanders recognised worst (see Table 2.12 for the WER breakdown per accent region). Table 2.4c lists the size of the bias against the five Dutch-speaking regions including Flanders, for every speaker age group, speech style, and ASR architecture, separately. Information in the brackets indicates which regions got the lowest and highest CER in every age group. All biases were shown to be significant ($p < .029$). For all age group,a bias in regional accents was observed. This finding is due to the fact that FL speakers were much worse recognised than any NL region's speakers regardless of age (see Table 2.12), which in turn is likely due to the lack of the use of FL training speech data.

In conclusion, both architectures showed clear biases against regional accents, particularly FL. This bias was similar for the hybrid system compared to the E2E system.

SUMMARY OF BIAS IN STATE-OF-THE-ART ASRS FOR DUTCH

The results showed that the Dutch ASR have 1) a gender bias, with a bias against male speech; 2) an age bias for native speakers of Dutch, with the largest bias against speech of native children, followed by speech of native older adults; 3) a bias against non-native speech, with an absolute WER degradation of around 24% in recognising non-native speakers' read speech and 15.0% in HMI speech; and 4) a bias against regional accents with the strongest biases against Flemish and speech from the south of the Netherlands. Comparing the biases exhibited by the two ASR architectures showed similar or smaller biases for the hybrid ASR system compared to the E2E system.

2.3.2. BIAS IN STATE-OF-THE-ART ASRS FOR MANDARIN

Unlike in the Dutch ASR experiments, there is no domain mismatch between training and test data in the Mandarin ASR experiments. The TDNNF hybrid system and the conformer E2E system achieved overall CER results of 3.3% and 2.9% respectively, on the MagicData test set.

BIAS AGAINST GENDER

Overall, female speech was recognised slightly better than male speech (see Table 2.13). Table (2.5a) lists the size of the bias against male Mandarin speakers for each of the regions separately. No significant differences between the CER for the female and male speakers, thus no bias, was observed for both the hybrid and the E2E models.

BIAS AGAINST REGIONAL ACCENTS

Overall, there is some variety in the recognition performance of the speech from the different accent regions, with speech from the Gan (GA) region being recognised best and that from the Min (MI) region recognised worst (see Table 2.13 for all WERs).

Table 2.5b lists the size of the bias against the various regions compared to the best-recognised region GA, separately for both genders. For the hybrid system, the largest bias occurred against speakers from Min (MI) and Xiang (XI) ($F(1,18)=14.165$, $p=.001$ and $F(1,18)=7.757$, $p=.012$) respectively). For the E2E system, only a bias against MI speech was observed ($F(1,18)=9.991$, $p=.005$).

In conclusion, both architectures showed a clear bias against MI (the worst recognised) speakers. Comparing the two ASR architectures shows that the E2E system was slightly less biased against regional (heavy) accents than the hybrid system, which also showed a bias against XI. Our finding regarding MI is in line with results reported in previous studies using a different database [27, 28].

SUMMARY OF BIAS IN STATE-OF-THE-ART ASRS FOR MANDARIN

In summary, the results showed that our SotA Mandarin ASRs showed no bias against gender. Regarding regional accents, our two ASR systems were both biased against MI speakers, and the hybrid system was also biased against XI speakers. The E2E ASR system was less biased against regional (heavy) accents than the hybrid system.

Table 2.5: Bias sizes for the TDNFF hybrid and conformer E2E ASR for gender (2.5a) and regional accents (2.5b) on MagicData. The region marked as † uses non-Mandarin Sinitic languages as local languages.

(a) Bias against gender, split by region. Female speech was recognised best.

Set	Hybrid		E2E	
	Size	P-value	Size	P-value
All	0.4	.419	0.4	.315
NG	-0.2	.593	-0.5	.377
SG	-0.1	.744	0.2	.634
†GA	-0.1	.761	0.2	.621
†MI	2.6	.054	2.2	.104
†WU	1.1	.197	1.3	.087
†XI	0.3	.827	0.0	.742
†YU	0.5	.744	1.1	.366

(b) Bias against regional accents. Bias sizes are calculated by subtracting the CER of GA from the CER of itself.

	Hybrid		E2E	
	Size	P-value	Size	P-value
NG	0.8	.176	0.7	.266
SG	0.4	.112	0.3	.467
†MI	2.7	.001	2.3	.005
†WU	0.2	.592	0.1	.919
†XI	1.0	.012	0.7	.081
†YU	1.1	.085	1.0	.098

2.4. FINDING THE ORIGIN OF BIAS

2.4.1. BIAS ACROSS ARCHITECTURES, SPEAKING STYLES, AND LANGUAGE

There are several important points that can be drawn from these results. First, bias and bias size are dependent on the architecture of the ASR system. For instance, we found a larger bias for the E2E models against male speakers for Dutch teenagers and older adults in both read speech and HMI speech, against non-native accents, against Flemish, and observed more bias against more strongly accented Mandarin.

Second, bias seems to be language-dependent. Although we can only compare bias against gender and region across Dutch and Mandarin, we do observe differences between the languages: while we found a bias against male speakers for Dutch for certain age groups, no gender bias was observed for the Mandarin speakers.

Third, bias was observed for both speaking styles, but seems to occur slightly more often for more spontaneous speech. Potentially HMI speech, which is less well prepared than read speech allows for more speaker-dependent articulations and differences in word usage, which cause an increase in recognition problems for the ASR systems.

2.4.2. PHONEME ANALYSIS INTO THE ORIGIN OF BIAS

In order to find the origin of bias, we focus our phoneme analysis on read speech as the bias seems to occur less often for read speech, so any results might also transfer across speaking styles. We compare across the two ASR architectures. We first identify

the phonemes that are worst recognised phonemes for those dimensions that showed a significant bias for each language individually, and then compare the phoneme error patterns found for Dutch and Mandarin in order to find common patterns.

2.4.3. DUTCH

Table 2.6 shows the breakdown of worst performing phonemes for the different dimensions (gender, age, non-nativeness, and native regional accents) for the two architectures separately. Blue colouring indicates a difference between the architectures; a red background colouring indicates a difference between speaker groups, for the specific phoneme.

The first thing to notice is the high similarity in the phonemes that are most difficult to recognise for each speaker group within each dimension: there are relatively few phonemes that are hardest to recognise that are not shared with the other speaker groups (not many phonemes with a red background). Also the relatively low number of blue colourings indicate that the architectures generally found the same phonemes hard to recognise. This is especially the case for gender, where four of the five worst recognised phonemes are shared. The bias against male speech can thus not be explained by a difference in pronunciation of specific phonemes which then would lead to specific phonemes being harder to recognise for male speech.

We observe a few more differences between the different age groups and between the two architectures. The hybrid system particularly seemed to have a problem recognising the /ə/, while the phoneme pattern of most difficult to recognise phonemes differs somewhat between the age groups. The observed bias against the age groups can thus partially be explained by differences in pronunciation of specific sounds.

For the non-native speakers, we find that the observed phonemes are those, which are known to be challenging to acquire for second language speakers, such as /œy/ and /y/, therefore pronunciation differences between native and non-native listeners seem to be a factor in the origin of the bias. This conclusion is also corroborated by increasing CEFs levels showing decreasing WER in the hybrid ASR architecture.

Across the regional accents, we mostly see differences between Flanders (FL) and NL (W, T, N and S), where /œy/ and /ɑu/ are among the most problematic phonemes for both architectures. The biases observed for the FL speakers thus likely have a pronunciation-based origin. We further see that /ɔ/ is a difficult sound to recognise in N(orth) regional accent.

2.4.4. MANDARIN

Table 2.7 lists the most problematic Mandarin phonemes for each regional accent. GA is the best recognised region whereas MI and XI are the worst recognised regions. It is clear that the bias against MI and XI cannot solely be explained by a difference in pronunciation of a range of phonemes: there is a large overlap in the phonemes that are worst recognised, except for the /s/. This latter finding can likely be explained by well-known variation patterns between the pronunciations of /ʂ/ and /s/, between /t̪ʂ/ and /ts/, and between /t̪ʂʰ/ and /tsʰ/ in Chinese regions using non-Mandarin Sinitic local languages (GA, MI, WU, XI and YU). We hypothesise that the high overall misrecognition of /s/ in MI and XI is caused by the /ʂ/ - /s/ ambiguity.

Table 2.6: The five worst performing phonemes for each dimension that was found to have a significant bias for the Dutch ASRs for both architectures. Blue colouring indicates a discrepancy between the hybrid and E2E architectures. Red background colouring indicates that the phoneme is only worst recognised for a particular speaker group.

	Hybrid					E2E				
	1	2	3	4	5	1	2	3	4	5
Gender										
Female	ʒ	ʃ	œy	ʏ	y	ʒ	ʃ	ʏ	y	œy
Male	ʃ	ʒ	œy	ʏ	ɲ	ʒ	ʃ	ɲ	œy	ʏ
Age										
DC	ʏ	h	ə	j	y	ʏ	f	y	h	b
DT	ʃ	h	ʏ	ə	j	ʏ	ʃ	h	œy	ɔ
DOA	h	ɔ	ə	ʏ	f	ʒ	h	x	ɔ	ʃ
Native and non-native accents										
AvgD	ʒ	ʃ	ʏ	h	ə	ʒ	ʃ	ʏ	h	f
AvgNN	œy	y	ʒ	ʏ	ɲ	ʒ	œy	y	ʏ	h
Regional accents										
W	ʃ	h	ʏ	ə	j	h	ʏ	ʒ	ʃ	ə
T	ʃ	ʒ	ʏ	h	ə	ʒ	ʃ	ʏ	f	h
N	ʃ	ʒ	h	ʏ	ɔ	ʒ	ʃ	ʏ	ɔ	h
S	ʒ	ʏ	h	ə	j	ʏ	h	ʒ	ə	f
FL	ʃ	ʒ	œy	ɑu	ɛi	ʒ	ʃ	œy	ɑu	ʏ

Table 2.7: The five worst performing phonemes for the accent regions for the Mandarin ASRs.

	Hybrid					E2E				
	1	2	3	4	5	1	2	3	4	5
Native regional accent										
GA	ʐ̩	ɛ	au	a	ə	ʐ̩	au	a	ei	s
MI	ʐ̩	s	ə	au	ei	ʐ̩	s	ə	au	l
XI	ʐ̩	s	ə	au	a	ʐ̩	s	au	l	ə

2.4.5. GENERAL PATTERNS

Comparing the different speaker groups within the different dimensions and across languages shows that most biases cannot solely be explained by pronunciation differences across multiple phonemes. Rather, the bias is likely due to differences at the supra-segmental level, i.e., the fundamental frequency (F0), which is much higher in children's voices than in adult's voices, speaking rate, which is typically slower in older adults than in younger adults, intonation, which differs substantially between Flemish and the Dutch Southern accent region and the other Dutch accent regions, and due to differences in word use.

Comparing the phoneme patterns across the architecture though shows that the hybrid and the E2E model show differences in which phonemes are hardest to recognise (blue phoneme symbols). In the case of the Dutch dataset, /j/ and /ə/ seem to be often poorly recognised by the hybrid system, while /f/ seems to be somewhat more difficult for the E2E system. For the Mandarin dataset, /l/ and /p^h/ are more problematic for the E2E system. There thus are phoneme-specific differences the E2E and Hybrid systems.

Finally, we observe that sounds that are less frequent in the language and thus the

training material (e.g., the /ʃ/, /ɲ/, /ʒ/ for Dutch), are typically less well recognised by the ASR systems. The composition of the training data thus plays an important role in the recognition of specific phonemes, and in creating and ultimately removing bias.

2.5. GENERAL DISCUSSION AND CONCLUSION

Our goal is to uncover bias in state-of-the-art DNN-based ASR systems to work towards proactive bias-mitigation in ASR systems in order to create inclusive automatic speech recognition for everyone, irrespective of how they speak or the language they speak. In this paper, we have focused on bias that can be quantified. However, owing to the foundational nature of bias, it is impossible to remove bias that creeps into datasets [29]. With this in mind, a priority in responsible ASR system development goes toward a proactive attitude. This concerns framing the problem, selecting the composition of the development team and the implementation process from a point of anticipating, proactively spotting, and developing mitigation strategies for prejudice.

A direct bias mitigation strategy concerns diversifying and aiming for a balanced representation of all types of speakers in the dataset [2, 17]. An indirect bias mitigation strategy deals with diverse team composition: the variety in age, regions, gender, etc. provides additional lenses of spotting potential bias in design. Together, they can help to ensure a more inclusive developmental environment for ASR.

In conclusion, there are big challenges to overcome before we reach the goal of significantly reducing the bias in ASR systems, and that these challenges are dependent on speaking style, language, and ASR architecture. This research shows that we should not focus on blindly lowering the error rates on our test sets but that it is crucial to take into account the speaker groups and demographics that are inherently present in our test set and, more importantly, in society.

2.6. DESIGN OF MANDARIN-SPEAKING REGIONS

We grouped speakers in the MagicData Mandarin speech corpus into seven accent regions based on the province-level geographical information that was provided in the corpus for all speakers. The province name(s) contained in each region are listed below:

- **Northern Guan (NG)**: Beijing, Gansu, Hebei, Heilongjiang, Henan, Jilin, Liaoning, Neimenggu, Ningxia, Shandong, Shanxi, Tianjin, Xinjiang;
- **Southern Guan (SG)**: Anhui, Chongqing, Jiangsu, Guizhou, Hubei, Sichuan, Yunnan;
- **Gan (GA)**: Jiangxi;
- **Min (MI)**: Fujian;
- **Wu (WU)**: Shanghai, Zhejiang;
- **Xiang (XI)**: Hunan;
- **Yue (YU)**: Guangdong, Guangxi.

Please note that in reality, more than one accents exist in some provinces. For instance, in Jiangsu, approximately 60% of the population uses SG, 30% of the population uses WU, and the rest uses NG [30]. We decided to label Jiangsu as SG, as the majority of the speakers use SG in that region.

2.7. IMPLEMENTATION DETAILS OF STATE-OF-THE-ART HYBRID AND E2E ASR SYSTEMS

2.7.1. HYBRID DNN-HMM ARCHITECTURE

The TDNN-BLSTM model for both the Dutch and Mandarin ASR systems consisted of three TDNN layers of dimension 1024, and 3 pairs of forward-backward LSTM layers of cell dimension 1024 on top. The TDNNF model for the Dutch and Mandarin ASRs consisted of 12 TDNNF layers of dimension 1024. For the Mandarin TDNNF model, we also added 6 convolutional layers between the input layer and the first TDNNF layer, following the recommended layout[7].

The language model (LM) in the hybrid ASR system is an RNNLM [25]. It consists of 3 TDNN layers interleaved with 2 LSTM layers. The RNNLM is trained with 20 epochs. To apply the RNNLM, a tri-gram LM is used to generate N-best results. After that, the RNNLM rescores the N-best results to get the final recognition results. The RNNLM and the tri-gram LM are trained using the training data transcriptions in CGN for Dutch and MagicData for Mandarin.

2.7.2. END-TO-END (E2E) ARCHITECTURE

The conformer E2E model parameters were mainly taken from [31, 32]: 12 encoder layers and 6 decoder layers, all with 2048 dimensions; the attention dimension is 512 and the number of attention heads is 8; the convolution subsampling layer in the encoder has 2-layer CNNs with 256 channels, stride with 2, and a kernel size of 3. The default kernel size (31) of the CONV module in the conformer structure was used for the Dutch ASR, while a CONV kernel size of 15 was used[8] for the Mandarin ASR. The conformer model was trained with 50 epochs using a joint connectionist temporal classification (CTC)-attention objective [33], in which the CTC and attention weights were set to 0.3 and 0.7, respectively. For the Dutch conformer model, subword units with a vocabulary size of 5000 were used as basic units. For the Mandarin conformer model, Chinese characters with a vocabulary size of 4481 were used as basic units.

An RNNLM was trained for each language, and used during E2E ASR decoding in a shallow fusion manner [26]. The RNNLM consisted of 2 LSTM layers of dimension 1024, and was trained with the training data transcripts of CGN (Dutch) or MagicData (Mandarin) for 40 epochs.

[7] `run_cnn_tdnn_1b.sh` in the Kaldi `multi_cn` recipe.
[8] This parameter is recommended in the ESPnet recipe of `aidatatang_200zh`.

Table 2.8: WERs of the TDNN-BLSTM and TDNNF hybrid ASRs and the conformer E2E ASR on the CGN standard broadcast news (BN) and conversational telephone speech (CTS) test sets. "F/M" indicates female/male. Numbers in bold indicate the best performance.

Arch.	Hybrid						E2E		
Model	TDNN-BLSTM			TDNNF			Conformer		
Set	Avg	F	M	Avg	F	M	Avg	F	M
BN	**5.6**	**5.5**	**5.6**	6.3	6.1	6.4	6.6	5.9	7.3
CTS	22.1	19.6	24.2	23.9	21.2	26.3	**21.6**	**18.8**	**24.0**

Table 2.9: Average WERs over different age groups in Jasmin-CGN native and non-native speakers' read speech by the TDNN-BLSTM and TDNNF hybrid ASR systems.

	TDNN-BLSTM	TDNNF
Native (average)	30.0	19.6
Non-native (average)	59.5	42.7

2.8. WORD ERROR RATE DETAILS OF THE DUTCH ASRS

2.8.1. IN-DOMAIN RESULTS

Table 2.8 lists the WER results on CGN (in-domain) test sets by the TDNN-BLSTM, TDNNF and conformer (E2E) systems.

2.8.2. OVERALL OUT-DOMAIN RESULTS

Table 2.9 compared TDNN-BLSTM and TDNNF systems on out-domain (Jasmin-CGN) test data.

2.8.3. WER BREAKDOWN FOR GENDER, AGE, NON-NATIVENESS AND REGIONAL ACCENTS

Table 2.10 shows the WER per age group, for the female and male speech separately and averaged over both genders (column Avg), for read speech and HMI separately. The top rows list the results for the native Dutch speakers per age group; the bottom rows for the non-native speakers per age group. The WERs per gender, averaged over all age groups (row Avg), over the native (row AvgD) and non-native (row AvgN) Dutch speakers, respectively, are also shown.

WERS FOR PER GENDER

Tables 2.10a and 2.10b show that, for both the hybrid and the E2E systems, in general, female speech was better recognised than male speech. Tables 2.10a and 2.10b also show that the average female-only and male-only read and HMI speech WERs achieved by the hybrid ASR system were all lower than the WERs by the E2E system.

Table 2.10: WERs of the TDNNF hybrid system (2.10a) and the conformer E2E system (2.10b) on the Jasmin-CGN read and HMI speech. "F/M" indicates female/male. "AvgD" indicates the average over all native Dutch speakers, "AvgN" over all non-native speakers and "Avg" indicates the average over all speakers.

(a) TDNNF hybrid system results

Group	Read			HMI		
	F	M	Avg	F	M	Avg
DC	25.6	25.9	25.8	31.5	28.5	30.0
DT	12.8	15.3	14.0	22.0	23.5	22.8
DOA	16.9	22.0	18.6	28.7	33.8	30.6
AvgD	18.3	21.1	19.6	28.4	30.8	29.4
NNC	41.5	42.6	42.0	42.0	43.0	42.5
NNA	43.4	43.8	43.6	42.2	45.9	43.7
AvgN	42.5	43.1	42.7	42.2	44.9	43.3
Avg	26.7	28.2	27.4	33.3	35.9	34.4

(b) Conformer E2E system results

Group	Read			HMI		
	F	M	Avg	F	M	Avg
DC	28.5	28.0	28.3	29.9	29.5	29.7
DT	14.5	17.7	16.1	20.6	24.3	22.5
DOA	18.3	23.6	20.0	28.2	34.7	30.7
AvgD	20.3	23.2	21.6	27.6	31.7	29.4
NNC	44.4	46.2	45.2	42.7	47.1	44.9
NNA	46.6	47.9	47.1	44.3	49.9	46.5
AvgN	45.5	46.9	46.1	43.9	49.0	46.1
Avg	29.1	30.8	29.8	33.4	38.0	35.3

WERS PER AGE GROUP

Tables 2.10a and 2.10b show that for both the hybrid and the E2E systems, among the native speakers, the Dutch teenager (DT) group achieved the best WER performances in read and HMI speech. Among the non-native speakers, the non-native children (NNC) group was slightly better recognised than the non-native adults (NNA) group. Comparing Table 2.10a with Table 2.10b shows that on both read and HMI speech, the hybrid system performed better than or similar to the E2E system in all the age groups.

WERS FOR NATIVE VS. NON-NATIVE SPEAKERS

Tables 2.10a and 2.10b show that speech of native speakers was recognised much better than that of non-native speakers of Dutch, regardless of speech types and ASR systems. Comparing Table 2.10a with Table 2.10b also shows that on both read and HMI speech, the hybrid system performed better than or similar to the E2E system on native and non-native speech.

Table 2.11 provides a closer look at the WERs for the different Dutch proficiency levels (CEF) of all the non-native adult speakers (NNA), separated by gender. It shows a reduction in the overall WER (column Avg) with an increase in CEF level (except WERs on read speech by the E2E system). This is in line with the intuition that non-native speakers with a higher CEF level tend to speak Dutch better than those with a lower CEF level.

Table 2.11: WERs of the TDNNF hybrid system (2.11a) and the conformer E2E system (2.11b) on the Jasmin-CGN non-native (NNA) speaker group separated by CEF proficiency in Dutch levels (A1 is the lowest level). "F/M" indicates female/male. The one B2-level speaker was omitted from the NNA speaker group for this analysis.

(a) TDNNF hybrid system results

CEF	Read			HMI		
	F	M	Avg	F	M	Avg
A1	44.6	44.4	44.5	43.7	47.6	47.0
A2	44.9	38.7	43.3	44.4	41.4	43.5
B1	37.6	51.5	42.6	38.4	44.7	40.4

(b) Conformer E2E system results

CEF	Read			HMI		
	F	M	Avg	F	M	Avg
A1	48.3	46.9	47.5	45.7	50.8	49.6
A2	46.4	44.5	45.9	46.1	46.2	46.1
B1	46.1	54.8	49.1	41.2	50.3	43.5

WERS PER NATIVE REGIONAL ACCENTS

Table 2.12 shows the WERs for each of the regional accents of the four large regions W, T, N and S in the Netherlands and Flanders (FL) in Belgium per age group, by the hybrid system (2.12a and 2.12b) and by the E2E system (2.12c and 2.12d). The average WER results over female and male speakers are shown in the grey rows, and the results broken down by female and male are shown in the white rows.

Table 2.12 shows that for both the hybrid and the E2E systems, speech spoken by people from Flanders (FL) achieved the worst performance in all age groups except for the older adults (DOA/FOA). Among the four regions in NL, for read speech, no region was consistently recognised worse than others; for HMI speech, region S in general was the worst recognised. Table 2.12 also shows that on read speech, the hybrid system performed better than the E2E system in all the four regions in NL and the region FL; on HMI speech, no superiority of one ASR system over the other was observed in the four regions in NL, while the hybrid system was found better than the E2E system in FL.

WERS FOR READ VS. HMI SPEECH

Table 2.10 shows that for both the hybrid system and the E2E system, the WER performance of HMI speech was much worse than that of read speech on native speaker groups. For the non-native speakers, the WER performances on read and HMI speech were very close for both ASR systems. The tiny performance gap between non-native speakers' HMI and read speech indicates that the clarity of articulation is different in native and non-native speakers – native speakers tend to enunciate while reading out loud and tend to articulate less well during spontaneous (HMI) speech, while this articulation difference due to speaking style seems to be less for non-native speakers.

2.9. CHARACTER ERROR RATE DETAILS OF THE MANDARIN ASRS

The CERs achieved by the TDNNF hybrid system and the conformer E2E system averaged over all speakers in the MagicData (adult-only) test set were 3.3% and 2.9% respec-

Table 2.12: WERs of the TDNNF hybrid system (2.12a and 2.12b) and the conformer E2E system (2.12c and 2.12d) on the Jasmin-CGN read and HMI speech of the four Dutch (NL) regions and Flanders in Belgium (BE) per age group. The average WERs are shown in the grey rows, and the WERs broken down by gender (female, male) are shown in the white rows.

(a) **Read** speech by TDNNF hybrid system

Country		NL			BE
Region	W	T	N	S	FL
DC/FC	N/A	23.8	28.3	25.6	35.3
	N/A	21.9,25.5	26.3,30.2	31.2,21.0	32.4,38.8
DT/FT	14.0	15.7	13.7	14.0	30.1
	12.7,15.0	13.2,17.7	12.8,14.6	12.6,15.5	28.6,31.8
DOA/FOA	17.2	19.0	13.3	25.0	22.5
	14.8,23.4	19.3,18.5	12.6,15.0	22.4,29.3	22.0,23.2

(b) **HMI** speech by TDNNF hybrid system

Country		NL			BE
Region	W	T	N	S	FL
DC/FC	N/A	31.4	27.0	30.1	47.5
	N/A	31.9,30.7	27.1,25.7	34.4,26.7	47.7,47.4
DT/FT	22.6	19.7	22.6	23.8	35.5
	19.1,25.8	19.2,19.9	23.1,21.6	23.4,23.9	34.6,36.7
DOA/FOA	29.0	29.3	24.3	37.4	36.4
	22.6,37.8	29.4,29.2	23.1,30.2	36.6,39.1	35.5,37.7

(c) **Read** speech by conformer E2E system

Country		NL			BE
Region	W	T	N	S	FL
DC/FC	N/A	26.5	30.9	27.7	39.1
	N/A	25.5,27.4	28.7,33.1	34.3,22.7	36.8,41.9
DT/FT	16.2	19.2	15.3	16.2	36.2
	14.2,17.9	16.0,22.3	14.2,16.3	14.8,17.9	33.3,39.5
DOA/FOA	18.7	20.2	14.7	26.9	24.3
	16.2,25.0	20.3,20.0	14.3,15.8	23.9,32.0	23.6,25.3

(d) **HMI** speech by conformer E2E system

Country		NL			BE
Region	W	T	N	S	FL
DC/FC	N/A	30.0	29.3	29.6	48.9
	N/A	29.7,30.4	28.2,30.9	32.4,27.4	48.5,49.3
DT/FT	20.4	19.9	22.7	25.0	39.8
	17.2,23.8	20.3,19.6	20.5,24.8	24.3,25.7	37.7,43.3
DOA/FOA	29.9	29.4	24.4	37.2	38.2
	24.1,37.9	29.4,29.3	22.4,33.7	35.1,41.4	37.9,38.6

Table 2.13: CERs of the TDNNF hybrid system and the conformer E2E system on the MagicData test sets. "F/M" indicates female/male. "Avg" indicates the average over all speakers. Regions with † use non-Mandarin Sinitic languages as their local languages.

Arch.	Hybrid			E2E		
Set	F	M	Avg	F	M	Avg
All	3.1	3.5	3.3	2.7	3.1	2.9
NG	3.3	3.1	3.2	3.0	2.5	2.9
SG	2.9	2.8	2.8	2.3	2.5	2.5
†GA	2.4	2.3	2.4	2.1	2.3	2.2
†MI	3.8	6.4	5.1	3.4	5.6	4.5
†WU	1.9	3.0	2.6	1.5	2.8	2.3
†XI	3.2	3.5	3.4	2.9	2.9	2.9
†YU	3.3	3.8	3.5	2.7	3.8	3.2

tively. Table 2.13 shows the CER for the female and male speech separately and averaged over both genders (column Avg), for every Mandarin accent region separately. The CERs per gender averaged over all the accent regions (row "All") are also shown.

2.9.1. CERs PER GENDER

Table 2.13 shows that, in general, female speech was better recognised than male speech. This is true for both the hybrid and the E2E architectures. This finding is in line with what has been found in the Dutch ASR experiments (see Section 2.8.3). Comparing the two ASR architectures shows that the E2E system outperformed the hybrid system on both female speech and male speech, both with an absolute CER reduction of 0.4%.

2.9.2. CERs PER REGIONAL ACCENT

Table 2.13 shows that the region GA achieved the best recognition performance among the seven Chinese accent regions, and this is true for both the hybrid and the E2E ASR architectures. The two regions using a variety of Mandarin as local languages, i.e., NG and SG, achieved CER results that were on par with or lower than the average CER over all the regions. This also means overall, the regions using non-Mandarin Sinitic languages as their local languages have higher CER than regions using Mandarin. The region MI, in which the local language is not Mandarin, had the worst CER results by both the hybrid and the E2E ASR systems. Comparing the two ASR architectures shows that the E2E system was consistently better than the hybrid system on every accent region.

BIBLIOGRAPHY

[1] M. Abu Shariah and M. Sawalha, "The effects of speakers' gender, age, and region on overall performance of arabic automatic speech recognition systems using the phonetically rich and balanced modern standard arabic speech corpus," in *Proceedings of the 2nd Workshop of Arabic Corpus Linguistics WACL-2*, 2013.

[2] A. Koenecke, A. Nam, E. Lake, J. Nudell, M. Quartey, Z. Mengesha, C. Toups, J. R. Rickford, D. Jurafsky, and S. Goel, "Racial disparities in automated speech recognition," *Proceedings of the National Academy of Sciences*, vol. 117, no. 14, pp. 7684–7689, 2020.

[3] M. Adda-Decker and L. Lamel, "Do speech recognizers prefer female speakers?" in *Proc. INTERSPEECH*, 2005.

[4] S. Goldwater, D. Jurafsky, and C. D. Manning, "Which words are hard to recognize? prosodic, lexical, and disfluency factors that increase speech recognition error rates," *Speech Communication*, vol. 52, no. 3, pp. 181–200, 2010.

[5] M. Garnerin, S. Rossato, and L. Besacier, "Gender representation in french broadcast corpora and its impact on asr performance," in *Proceedings of the 1st International Workshop on AI for Smart TV Content Production, Access and Delivery*, 2019, pp. 3–9.

[6] R. Tatman, "Gender and dialect bias in youtube's automatic captions," in *Proceedings of the First ACL Workshop on Ethics in Natural Language Processing*, 2017, pp. 53–59.

[7] R. Tatman and C. Kasten, "Effects of talker dialect, gender & race on accuracy of bing speech and youtube automatic captions." in *Proc. INTERSPEECH*, 2017, pp. 934–938.

[8] Y. Qian, K. Evanini, X. Wang, C. M. Lee, and M. Mulholland, "Bidirectional lstm-rnn for improving automated assessment of non-native children's speech." in *INTERSPEECH*, 2017, pp. 1417–1421.

[9] L. Moro-Velázquez, J. Cho, S. Watanabe, M. A. Hasegawa-Johnson, O. Scharenborg, H. Kim, and N. Dehak, "Study of the performance of automatic speech recognition systems in speakers with parkinson's disease," in *Proc. INTERSPEECH*, 2019, pp. 3875–3879.

[10] B. M. Halpern, R. van Son, M. W. M. van den Brekel, and O. Scharenborg, "Detecting and analysing spontaneous oral cancer speech in the wild," in *Proc. INTERSPEECH*, 2020, pp. 4826–4830.

[11] M. Schuster, A. Maier, T. Haderlein, E. Nkenke, U. Wohlleben, F. Rosanowski, U. Eysholdt, and E. Nöth, "Evaluation of speech intelligibility for children with cleft lip and palate by means of automatic speech recognition," *International Journal of Pediatric Otorhinolaryngology*, vol. 70, no. 10, pp. 1741–1747, 2006.

[12] Y. Wu, D. Rough, A. Bleakley, J. Edwards, O. Cooney, P. R. Doyle, L. Clark, and B. R. Cowan, "See what i'm saying? comparing intelligent personal assistant use for native and non-native language speakers," in *22nd International Conference on Human-Computer Interaction with Mobile Devices and Services*, 2020, pp. 1–9.

[13] A. Palanica, A. Thommandram, A. Lee, M. Li, and Y. Fossat, "Do you understand the words that are comin outta my mouth? voice assistant comprehension of medication names," *NPJ digital medicine*, vol. 2, no. 1, pp. 1–6, 2019.

[14] C. Cucchiarini, O. van Herwijnen, F. Smits *et al.*, "JASMIN-CGN: Extension of the spoken dutch corpus with speech of elderly people, children and non-natives in the human-machine interaction modality," in *Proc. LREC*, 2006.

[15] G. I. Winata, S. Cahyawijaya, Z. Liu, Z. Lin, A. Madotto, P. Xu, and P. Fung, "Learning fast adaptation on cross-accented speech recognition," in *Interspeech 2020, 21st Annual Conference of the International Speech Communication Association, Virtual Event, Shanghai, China, 25-29 October 2020*, H. Meng, B. Xu, and T. F. Zheng, Eds. ISCA, 2020, pp. 1276–1280. [Online]. Available: https://doi.org/10.21437/Interspeech.2020-0045

[16] E. Alsharhan and A. Ramsay, "Investigating the effects of gender, dialect, and training size on the performance of arabic speech recognition," *Language Resources and Evaluation*, vol. 54, no. 4, pp. 975–998, 2020.

[17] A. Caliskan, J. J. Bryson, and A. Narayanan, "Semantics derived automatically from language corpora contain human-like biases," *Science*, vol. 356, no. 6334, pp. 183–186, 2017.

[18] N. Oostdijk, "The spoken Dutch corpus. overview and first evaluation." in *LREC*. Athens, Greece, 2000, pp. 887–894.

[19] D. A. v. Leeuwen, J. Kessens, E. Sanders, and H. v. d. Heuvel, "Results of the n-best 2008 dutch speech recognition evaluation," in *Proc. INTERSPEECH*, 2009.

[20] Magic Data Technology Co., Ltd., "MAGICDATA Mandarin Chinese Read Speech Corpus," 2019. [Online]. Available: http://www.imagicdatatech.com/index.php/home/dataopensource/data_info/id/101

[21] D. Povey, G. Cheng, Y. Wang, K. Li, H. Xu, M. Yarmohammadi, and S. Khudanpur, "Semi-orthogonal low-rank matrix factorization for deep neural networks," in *Proc. INTERSPEECH 2018*, 2018, pp. 3743–3747.

[22] D. Povey, A. Ghoshal, G. Boulianne, L. Burget, O. Glembek, N. Goel, M. Hannemann, P. Motlicek, Y. Qian, P. Schwarz *et al.*, "The Kaldi speech recognition toolkit," in *Proc. ASRU*, 2011.

[23] S. Feng and T. Lee, "Improving cross-lingual knowledge transferability using multilingual TDNN-BLSTM with language-dependent pre-final layer," in *Proc. INTERSPEECH*, 2018, pp. 2439–2443.

[24] S. Watanabe, T. Hori, S. Karita, T. Hayashi, J. Nishitoba, Y. Unno, N. Enrique Yalta Soplin, J. Heymann, M. Wiesner, N. Chen, A. Renduchintala, and T. Ochiai, "Espnet: End-to-end speech processing toolkit," in *Proc. INTERSPEECH*, 2018, pp. 2207–2211.

[25] H. Xu, K. Li, Y. Wang, J. Wang, S. Kang, X. Chen, D. Povey, and S. Khudanpur, "Neural network language modeling with letter-based features and importance sampling," in *Proc. ICASSP*, 2018, pp. 6109–6113.

[26] T. Hori, S. Watanabe, Y. Zhang, and W. Chan, "Advances in joint CTC-attention based end-to-end speech recognition with a deep CNN encoder and RNN-LM," in *Proc. INTERSPEECH*, 2017, pp. 949–953.

[27] J. Yi, Z. Wen, J. Tao, H. Ni, and B. Liu, "CTC regularized model adaptation for improving LSTM RNN based multi-accent mandarin speech recognition," *J. Signal Process. Syst.*, vol. 90, no. 7, pp. 985–997, 2018.

[28] H. Zheng, S. Zhang, L. Qiao, J. Li, and W. Liu, "Improving large vocabulary accented mandarin speech recognition with attribute-based i-vectors," in *Proc. INTERSPEECH*, 2016, pp. 3454–3458.

[29] O. Kudina and B. de Boer, "Co-designing diagnosis: Towards a responsible integration of machine learning decision-support systems in medical diagnostics," *Journal of Evaluation in Clinical Practice*, 2021.

[30] W. contributors, "Jiangsu province," 2021, [Online; accessed 21-July-2021]. [Online]. Available: https://zh.wikipedia.org/w/index.php?title=%E6%B1%9F%E8%8B%8F%E7%9C%81&oldid=66464253

[31] P. Guo, F. Boyer, X. Chang, T. Hayashi, Y. Higuchi, H. Inaguma, N. Kamo, C. Li, D. Garcia-Romero, J. Shi *et al.*, "Recent developments on espnet toolkit boosted by conformer," in *Proc. ICASSP*, 2021, pp. 5874–5878.

[32] S. Karita, X. Wang, S. Watanabe, T. Yoshimura, W. Zhang, N. Chen, T. Hayashi, T. Hori, H. Inaguma, Z. Jiang, M. Someki, N. E. Y. Soplin, and R. Yamamoto, "A comparative study on transformer vs RNN in speech applications," in *Proc. ASRU*, pp. 449–456.

[33] S. Kim, T. Hori, and S. Watanabe, "Joint ctc-attention based end-to-end speech recognition using multi-task learning," in *2017 IEEE international conference on acoustics, speech and signal processing (ICASSP)*. IEEE, 2017, pp. 4835–4839.

3

Low-Resource Automatic Speech Recognition and Error Analyses of Oral Cancer Speech

In this paper, we introduce a new corpus of oral cancer speech and present our study on the automatic recognition and analysis of oral cancer speech. A two-hour English oral cancer speech dataset is collected from YouTube. Formulated as a low-resource oral cancer ASR task, we investigate three acoustic modelling approaches that previously have worked well with low-resource scenarios using two different architectures; a hybrid architecture and a transformer-based end-to-end (E2E) model: (1) a retraining approach; (2) a speaker adaptation approach; and (3) a disentangled representation learning approach (only using the hybrid architecture). The approaches achieve a (1) 4.7% (hybrid) and 7.5% (E2E); (2) 7.7%; and (3) 2.0% absolute word error rate reduction, respectively, compared to a baseline system which is not trained on oral cancer speech. A detailed analysis of the speech recognition results shows that (1) plosives and certain vowels are the most difficult sounds to recognise in oral cancer speech - this problem is successfully alleviated by our proposed approaches; (3) however these sounds are also relatively poorly recognised in the case of healthy speech with the exception of /p/. (2) recognition performance of certain phonemes is strongly data-dependent; (4) In terms of the manner of articulation, E2E performs better with the exception of vowels - however, vowels have a large contribution to overall performance. As for the place of articulation, vowels, labiodentals, dentals and glottals are better captured by hybrid models, E2E is better on bilabial, alveolar, postalveolar, palatal and velar information. (5) Finally, our analysis provides some guidelines for selecting words that

Appeared as: Halpern, B. M., Feng, S., van Son, R., van den Brekel, M., & Scharenborg, O. (2022). Low-resource automatic speech recognition and error analyses of oral cancer speech. Speech Communication. The PhD candidate contributed to the data collection, experimental design, writing, and evaluation of the experiments.

can be used as voice commands for ASR systems for oral cancer speakers.

3.1. INTRODUCTION

It is a great problem that many assistive technologies are only accessible to people with unimpaired speech. Often those who have the biggest need of such technologies are deprived of them. Oral cancer survivors are one such group of speakers. Approximately 500,000 people get diagnosed with oral cancer every year worldwide [1], of which 53,000 in the USA [2] alone.

Oral cancer leads to speech impairments due to the (partial) removal of the tissues surrounding the tongue during surgery as part of the treatment of the oral cancer [3]. Oral cancer speakers' speech impairments are predominantly on the articulatory level. Plosives (i.e. /k/, /g/, /b/, /p/, /t/, /d/) [4, 5] and alveolar sibilants (i.e., /s/, /z/) [6] have been found to be the most impacted [7]. In certain cases, patients are able to learn articulatory compensation techniques to adjust for the lost tongue tissue [3]. Their impaired ability to speak affects their quality of life to a great extent [8]. This comes in addition to difficulty swallowing, chewing [3, 9], and decreased tongue mobility [10] after operation.

This research focuses on building an automatic speech recognition (ASR) system for oral cancer speech. Such an ASR could have a large positive impact on survivors' quality of life and could be used in the objective evaluation of survivors' speech intelligibility during speech therapy [11]. To that end, this paper 1) presents a newly collected database of English oral cancer speech; 2) investigates several approaches to building an ASR for oral cancer speech, where we specifically focus on the acoustic model to improve oral cancer speech recognition (and leave sophisticated language models and data augmentation for future research; see also the General Discussion); and 3) presents an analysis into the differences and similarities between oral cancer speech and normal speech.

Training a deep neural network (DNN) acoustic model (AM) for the automatic recognition of speech usually requires a large amount of labelled training data. In the case of oral cancer speech, though, we typically only have a very limited amount of labelled oral cancer speech data. This makes DNN AM training for oral cancer speech a low-resource problem. We investigate three hybrid approaches in low-resource ASR that previously have been shown to be competitive on low-resource tasks: (1) a retraining approach [12], (2) a speaker adaptation approach [13], and (3) a disentangled representation learning approach [14] in order to leverage non-pathological, normal speech resources in DNN AM training for building AMs for oral cancer speech. (4) Due to the recent success of end-to-end (E2E) architectures, we additionally perform DNN AM retraining with a Transformer-based ASR architecture.

The acoustic model retraining approach leverages an AM pretrained on a healthy speech corpus and retrains this AM with oral cancer speech data. This approach has shown to be effective in improving acoustic modelling for pathological speech [15, 16], including dysarthria [17, 18], and aphasia [19] for hybrid models. An effective multi-stage acoustic modelling method for dysarthric speech was proposed in [17].

Transformer-based E2E models are known to perform well when exposed to a large amount of training data and for standard, general-purpose ASR tasks [20]. There is, however, limited research in pathological ASR using a Transformer-based architecture,

with the exception of [21] for dysarthric ASR. However, the Transformer-based model achieves worse WER performance (even with data augmentation) than the current state-of-the-art [18]. The present study adopts a similar method to [17], and studies the efficacy of the retraining approach for the recognition of oral cancer speech using a hybrid and an E2E model.

The goal of speaker adaptation, or speaker adaptive training (SAT), is to normalise speaker variation contained in speech [22], and is widely applied in general-purpose ASR systems [23, 24, 25, 26]. It is expected that speaker adaptation is even more important in oral cancer ASR, as oral cancer speech is much more variable than normal speech. We propose to use speaker adaptation, and particularly feature-space maximum likelihood linear regression (fMLLR) [27] based speaker adaptation, to suppress pathological speech sound characteristics in oral cancer speech, encouraging oral cancer speech representations to be more similar to those of normal speech. fMLLR has previously been successfully applied to improve pathological speech recognition performance [28, 16, 29]. The resulting AM is expected to perform better on the oral cancer ASR task than that without speaker adaptation.

Disentangled speech representation learning aims to separate phonetic and speaker information in the speech signal into two feature representations in an unsupervised manner [14], i.e., without the need of labelled speech data. One of the two learned representations, the phonetically-discriminative representation, is expected to retain the linguistic content in the original speech signal while suppressing speaker-dependent information. Conversely, the other learned representation is expected to capture speaker-dependent information and carry little phonetic information. The effectiveness of disentangled representation learning has been demonstrated for low-resource ASR [30, 31] and noise robust ASR [32]. In the present study, we propose to apply this approach to suppress pathological speech sound characteristics while retaining the linguistic content in the oral cancer speech. Specifically, we adopt the factorised hierarchical variational auto-encoder (FHVAE) [14] to perform disentangled speech representation learning. The learned phonetically-discriminative feature representation is used as the input feature to train a DNN AM for the oral cancer ASR task.

We further carry out an extensive phoneme-level and articulatory-level analysis in Section 3.4.2. The goal of this analysis is five-fold:

- Firstly, we want to find out what phonemes and articulatory features of the oral cancer speech are the most difficult to capture for current ASR systems trained on typical speech. This will allow us to investigate whether these sounds are the sounds that are known to be impacted in oral cancer speech or if ASR systems have problems with other sounds or aspects of oral cancer speech.

- Secondly, we want to pinpoint which phonemes and articulatory features contribute most to improvements in the proposed ASR systems. The motivation for this analysis is to identify performance bottlenecks, which will guide the development of future ASR systems. It is especially important to pinpoint phoneme classes where adding more oral cancer speech data is not expected to help. We would like to see which phonemes are better recognised by E2E models/hybrid models in the case of oral cancer speech. End-to-end models became superior to hybrid models

on many ASR tasks, therefore we hypothesise that for certain sounds end-to-end models will be better. Determining which ones are better are essential for choosing the appropriate architecture for future pathological speech studies.

- Thirdly, we would like to compare the errors that the ASR architectures make on healthy and oral cancer speech. The goal of this analysis is to pinpoint which phoneme classes are specific to oral cancer speech, and which phonemes seem to be problematic for both kind of speech.
- Fourthly, the outcomes of the analyses will be used to provide guidelines on the selection of the words used for voice commands or stimuli for ASR systems aimed at oral cancer speakers. For example, if a particular class of phonemes are better recognised by the proposed systems than other phonemes, a voice command consisting mostly of phonemes from that class of phonemes can be selected. Such an analysis could bear meaningful lessons when deploying these systems to voice assistant tools or when these are used for objective evaluation of oral cancer speech.

Finally, it is well known that background noise negatively affects the performance of ASR systems [33]. Our dataset was collected from YouTube, which left us with little control regarding the noise in the audio. Therefore, it would be useful to quantify the influence of noise on the ASR performance, and compare it to the influence of speech severity. In Section 3.3.6, we perform an analysis to compare the influence of noise and speech severity in our ASR systems.

3.2. DATASET

In our experiments, we will use two datasets: a new, publicly available dataset we have recently collected containing English oral cancer speech[1]; and the Wall Street Journal (WSJ) dataset [34] containing English (*non-pathological*) read speech to leverage as training data for our baseline system and as a starting point for training our low-resource scenario ASR systems.

3.2.1. ORAL CANCER SPEECH DATASET

We manually collected 2.25 hours of audio data containing English oral cancer speech from 10 different speakers from YouTube. Presence of oral cancer speech was determined by the content of the video and the authors' (B.H., R.v.S, M.v.d.B) clinical experience with such speakers. The audio was then manually cut to exclude music, healthy speakers, non-American English speakers, unintelligible speech, and other factors which could negatively influence recognition of the oral cancer speech. The resulting corpus has been automatically cut into chunks of 10 second. The cuts do not necessarily occur at natural pauses. When we transcribed the utterances, we tried to account for this as much as possible.

Baseline transcriptions were generated using the *Baseline* ASR system used in this study, which consisted of a DNN AM and a tri-gram language model (LM; see Section 3.3.1). Subsequently, these automatic transcriptions were manually checked and corrected by one of the authors (B.H.).

[1] https://karkirowle.github.io/oral_cancer_corpus/

Table 3.1 shows the number of recordings and the amount of speech in minutes for each of the recordings of each of the speakers, as well as the speakers' gender. Since the total amount of oral cancer speech data is rather limited and because the total amount of audio for each speaker is highly variable, we carried out 5-fold cross-validation rather than creating separate training and test sets.

A completely random, blind shuffling of the speakers into the five separate training and test sets would lead to (1) high variance in the observed WERs due to the large differences in the amount of audio used for training and testing in each possible partition, (2) high gender imbalance, i.e., in a completely random shuffling, an all-male train and all-female test set could easily occur. Therefore, to create the five training-test set combinations, the train and test set speakers are selected so that (1) the total audio used for training is always around 100 mins (1.7 hours), and (2) the gender balance of the train/test set varies within acceptable ranges, so that the training set contains at least two speakers of the same gender; and at least one speaker of that same gender is present in the test set. As a large portion of the audio data comes from the speaker with ID id011 (see Table 1), this speaker is always kept in the training set. The partitions are shown in Table 3.1. The speakers are either assigned to the training set or the test set, there is no overlap. The amounts of audio data in hours, the total numbers of words in the transcriptions, and the total number of audio files in the training and test data separated per gender are listed in Table 3.2 for each partition separately.

3.2.2. WALL STREET JOURNAL CORPUS

The Wall Street Journal (WSJ) corpus is an American English read speech corpus [34]. We used the `si284` set, which contains 37,416 speech utterances spoken by 283 speakers, for training. The total amount of data is 81.3 hours. All speakers in the WSJ are healthy speakers.

3.3. METHODS

The three approaches with the two different architectures to the automatic recognition of oral cancer speech will be compared against two *Baseline* ASR systems - one hybrid system and one E2E system - on the task of word recognition on the oral cancer speech test set. Word recognition performance is measured in word error rate (WER). We also report WER on the oral cancer speech training set, which is used in the analyses of the oral cancer recognition results (see Section 3.4.1). Figures 3.1 present a schematic overview of the three approaches and the *Baseline* model implemented in the hybrid DNN-HMM architecture (top of Figure 3.1). For ease of comparison of the three approaches, we used colours to indicate similarities (and differences) between the approaches: The blue colour indicates GMM-HMM training, the green colour indicates DNN AM (re-)training (the same approach is used for both architectures), and the orange colour indicates the feature representation method (only for the hybrid approach). The dashed boxes indicate the type of data that is used in the various stages of the pipelines of the three approaches.

An overview of the training data, feature representations, and training methods of the three approaches and the *Baseline* models implemented in the hybrid and E2E ar-

Table 3.1: Details of the oral cancer speech dataset and its train-test partitioning design for 5-fold cross-validation. **Blue** means train, while **red** means test.

Wav id	Spk id	Minutes	Gender	Partition index				
				1	2	3	4	5
1	id001	1.6	female	test	test	train	train	train
3		3.3		test	test	train	train	train
10	id003	17.5	female	train	train	train	test	train
21	id007	12.8	female	train	train	train	train	test
23	id008	6.2	female	train	test	test	train	train
24		15.0		train	test	test	train	train
18	id005	6.1	female	test	test	test	train	train
4	id011	1.4	male	train	train	train	train	train
5		4.2		train	train	train	train	train
6		2.9		train	train	train	train	train
7		3.2		train	train	train	train	train
13		4.1		train	train	train	train	train
22		11.9		train	train	train	train	train
28		13.9		train	train	train	train	train
26	id011/id009	13.3	mixed	train	train	train	train	train
30	id014	0.4	male	test	test	test	train	train
33	id016	1.8	male	test	test	test	test	train
34	id017	15.5	male	test	train	train	test	test

Table 3.2: Statistics of the training and test data in the 5-fold cross-validation scheme.

Partition index		1	2	3	4	5
Training set	Hours	1.77	1.68	1.76	1.67	1.78
	#words	17.2k	16.7k	17.3k	17.2k	17.5k
	#male audio files	7	8	8	8	9
	#female audio files	4	2	4	6	6
	#mixed audio files	1	1	1	1	1
Test set	Hours	0.48	0.57	0.49	0.58	0.47
	#words	4.7k	5.3k	4.6k	4.7k	4.4k
	#male audio files	3	5	3	1	1
	#female audio files	3	2	2	2	1
	#mixed audio files	0	0	0	0	0

Table 3.3: Attributes of the hybrid and E2E models compared in this study. FB+P: FBank + Pitch feature. OC: oral cancer speech. '→': pretraining followed by retraining. '+': merging of the two datasets during training.

Architecture	Hybrid				E2E	
Attributes / Method	GMM-HMM		DNN		Transformer	
	Data	Input	Data	Input	Data	Input
Baseline	WSJ	MFCC	WSJ	FB+P	WSJ	FB+P
DNN AM/E2E ASR retraining	WSJ	MFCC	WSJ→OC	FB+P	WSJ→OC	FB+P
Baseline+OC	WSJ+OC	MFCC	WSJ+OC	FB+P	N/A	
fMLLR for AM training	WSJ+OC	MFCC	WSJ+OC	fMLLR	N/A	
FHVAE	WSJ+OC	MFCC	WSJ+OC	$\boldsymbol{z}_1$	N/A	

chitectures is provided in Table 3.3.

3.3.1. *Baseline* ASR SYSTEMS

BASELINE HYBRID ASR

The *Baseline* hybrid ASR system is visualised in the left part of Figure 3.1 (top) in the part of the pipeline that says "WSJ data", and consists of a hybrid DNN-hidden Markov model (DNN-HMM) AM only trained with WSJ. The input features of the *Baseline* system are 23-dimension filter banks (FBanks) appended by 3-dimension pitch features [35]. The 26-dimension features are further processed by contextual splicing $\{0, \pm1, \pm2, \pm3, \pm4, \pm5\}$ (following the recommendation in Kaldi[2]), i.e., each frame-level feature appended by its 5 left and 5 right frames, to capture longer temporal dependencies. This results in 286 $(26 \times (5+1+5))$ dimensions.

To obtain the phone labels for each speech frame of the WSJ data for DNN training, first a context-dependent GMM-HMM (CD-GMM-HMM) AM is trained from scratch with the WSJ training data and transcriptions using the standard Kaldi recipe [36]. The CMU dictionary[3] is used to map the words in the training data transcriptions to sequences of phonemes. The input features are 39-dimension MFCCs+Δ+$\Delta\Delta$. After CD-GMM-HMM AM training, the number of modelled HMM states is 3,431. Frame labels are then obtained via forced alignment with the CD-GMM-HMM.

The DNN contains 5 feed-forward layers of dimension 1,500 and a softmax output layer of dimension 3,431 (equal to the number of HMM states). The DNN AM is trained using the WSJ frame labels as training labels and cross-entropy (CE) [37] as the training criterion, and implemented based on Kaldi `nnet1`[4]. A 10% subset of training data is randomly selected for cross-validation (CV). The initial learning rate (LR) is 0.008, and is halved when no improvement of the loss value in the CV set is observed. Following the Kaldi `nnet1` convention, the training process is terminated if the LR is smaller than 1.5625×10^{-5}.

[2] `wsj/s5/steps/nnet/pretrain_dbn.sh`

[3] `http://www.speech.cs.cmu.edu/cgi-bin/cmudict`

[4] The LF-MMI criterion [38] was found more effective than CE in dysarthric ASR [18]. However, our initial experiments using DNN AM trained with LF-MMI using the more recent `nnet3` in Kaldi showed no improvements over training using CE.

The *Baseline* ASR system uses a tri-gram LM trained with the transcriptions of the WSJ `si284` set. This LM is adopted consistently throughout all experiments in this paper[5]. The *Baseline* ASR system achieves a WER of 6.7%[6] on the official WSJ test set `eval92`.

BASELINE END-TO-END (E2E) ASR

The *baseline E2E* ASR system adopts a transformer architecture [20] and is, like the Baseline hybrid model, only trained with the WSJ training material. The input features of the *E2E Baseline* system are 23-dimension FBanks appended by 3-dimension pitch features, the same input features as used for the *Baseline hybrid* system as described in Section 3.3.1. The transformer model parameters are taken mainly from the official ESPnet WSJ recipe[7]: 12 encoder layers and 6 decoder layers, all with 2048 dimensions; the attention dimension is 256 and the number of attention heads is 4; the convolution subsampling layer in the encoder has 2-layer CNN with 256 channels, stride with 2, and a kernel size of 3. The transformer model is trained with 50 epochs (no early-stopping), with a LR of 10.0, using a joint connectionist temporal classification (CTC)-attention objective [40] in which the CTC and attention weights are 0.3 and 0.7 respectively. Letters of the English alphabet are used as the basic subword units. The *E2E baseline* achieved 5.3% WER on the WSJ eval92 test set.

3.3.2. MODEL RETRAINING

In this approach, an ASR system is first trained with normal, i.e., in this case WSJ, speech data, and then retrained with oral cancer speech data. Sections 3.3.2 and 3.3.2 discuss the retraining approach applied to the hybrid and E2E ASR architectures, respectively.

HYBRID DNN AM RETRAINING

The general framework of applying the retraining approach to a hybrid ASR system is illustrated in the top part of Figure 3.1. The *Baseline* DNN AM described in Section 3.3.1 is chosen as the pretrained model and used as the starting point for retraining.

First, the *Baseline* DNN AM is used to force-align the oral cancer speech. Then, these alignments are used as labels to retrain the *Baseline* model. Preliminary experiments compared retraining some of the hidden layers vs. all hidden layers. The results showed that retraining all the hidden layers gave the best WER on the oral cancer speech test set. Therefore, *DNN AM retraining* in this study is always performed on all the hidden layers.

The loss function and stopping criterion of *DNN AM retraining* are the same as those for the *Baseline* DNN AM training. The initial LR was carefully tuned using the oral cancer speech data of partition 1 in the range of $\{0.002, 0.004, 0.008, 0.016\}$ because we discovered that with very limited amounts of oral cancer training speech data for *DNN AM retraining*, the WER performance on the oral cancer speech was sensitive to the initial LR. Our preliminary experiments showed that the optimal LR was 0.008, and it is used in all experiments in this paper.

[5] RNNLM rescoring on top of tri-gram LM based results could lead to a WER reduction, however, this paper focuses on acoustic modelling, hence RNNLM rescoring to a hybrid model is not applied in this paper.

[6] This result falls short of state of the art [39], mainly due to (1) the use of a tri-gram LM, and (2) the use of CMUdict without the extension to include the out-of-vocabulary words in the WSJ LM training data.

[7] `egs/wsj/asr1/conf/tuning/train_pytorch_transformer.yaml` from ESPNet

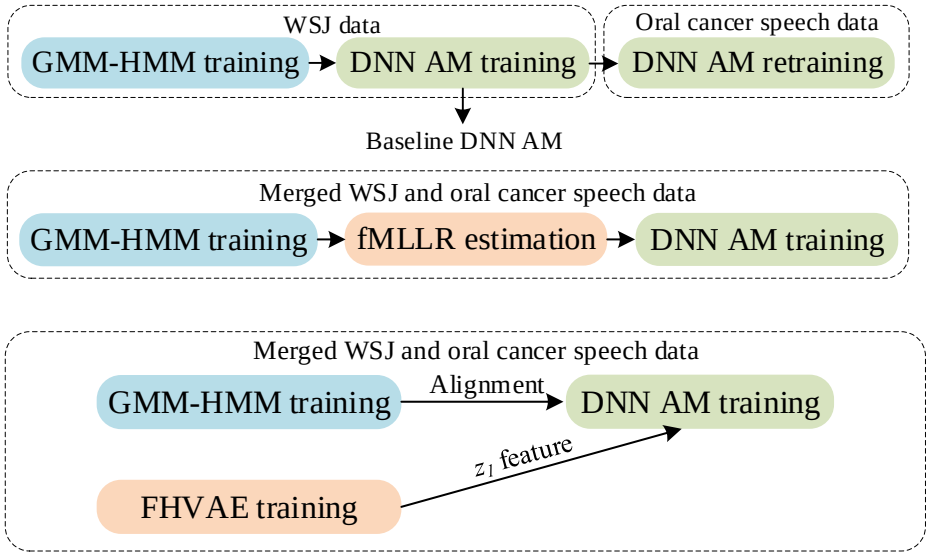


Figure 3.1: (Top) Schematic overview of the *DNN AM retraining* approach. The left-most part, indicated with the dashed lines, shows the *Baseline* model. (Middle) Schematic overview of the *fMLLR* for AM training approach. (Bottom) Schematic overview of the disentangled speech representation learning for AM training approach.

E2E ASR RETRAINING

The *baseline E2E ASR* system described in Section 3.3.1 is chosen as the pretrained model and used as the starting point for retraining. We carried out *E2E ASR retraining* on all the encoder and decoder network layers of the pretrained model, in order to be consistent with the setup in *hybrid DNN AM retraining* (see Section 3.3.2). Similarly, as in the case of *hybrid DNN AM retraining*, we experimentally found the performance of *E2E ASR retraining* is sensitive to the LR. Our results indicated that the optimal LR for transformer is 0.5, and it is used in all *E2E ASR retraining* experiments in this paper.

3.3.3. SPEAKER-ADAPTED FEATURES FOR ACOUSTIC MODELLING

The idea of fMLLR is to map acoustic speech features from the original unadapted space to a speaker-adapted space, so that the adapted features are less dependent on speaker identities. This is realised by the estimation of speaker-specific transform matrices and bias vectors.

Mathematically, let $\boldsymbol{o}_t^s$ be an unadapted speech feature at frame t, spoken by speaker s. fMLLR estimates a matrix $\mathbf{A}^s$ and a bias vector $\boldsymbol{b}^s$, and transforms $\boldsymbol{o}_t^s$ to $\hat{\boldsymbol{o}}_t^s$ by,

$$\hat{\boldsymbol{o}}_t^s = \mathbf{A}^s \cdot \boldsymbol{o}_t^s + \boldsymbol{b}^s, \tag{3.1}$$

where $\hat{\boldsymbol{o}}_t^s$ is the corresponding speaker adapted feature. The estimation of $\mathbf{A}^s$ and $\boldsymbol{b}^s$ can be realised by an expectation-maximisation (EM) algorithm proposed in [27]. The speaker adapted features $\hat{\boldsymbol{o}}_t^s$ are also often referred to as fMLLR features.

The use of fMLLR features in acoustic modelling for oral cancer speech is illustrated in Figure 3.1 (middle). The oral cancer speech data and WSJ data are merged to train a CD-GMM-HMM AM from scratch using the training procedure of the *Baseline* ASR system (see Section 3.3.1), except that here we also include the oral cancer speech data in the training of the CD-GMM-HMM AM model. Subsequently, fMLLR-based SAT is performed on the CD-GMM-HMM AM to estimate speaker-specific matrices and bias vectors. After SAT, a new CD-GMM-HMM AM with fMLLR features as input features is trained. This model is denoted as the CD-GMM-HMM-SAT. The dimension of fMLLR features is 40. The number of HMM states modelled by the CD-GMM-HMM-SAT model is 5,080. Next, frame alignments are generated with CD-GMM-HMM-SAT for both the WSJ and the oral cancer speech data. These alignments and fMLLR features are used as training labels and input features, respectively, to train a DNN AM for oral cancer ASR.

In short, the DNN training procedure and architecture follow the settings of the *Baseline hybrid* DNN AM training, except: (1) Training data consists of both WSJ and oral cancer speech; (2) The softmax output layer dimension is 5,080; (3) Input features to the DNN AM are fMLLR features, instead of FBank+pitch features. This method is denoted as *fMLLR for AM training* (or *fMLLR* for simplicity), and is only carried out for the hybrid architecture.

To explicitly measure the efficacy of fMLLR-based speaker adaptation, we trained another DNN AM, which takes 23-dimension FBanks appended by 3-dimension pitch features as input, instead of fMLLR features. Other training and model settings are the same as the system with *fMLLR for AM training*. This system is referred to as *Baseline+OC*, where OC stands for oral cancer speech.

3.3.4. DISENTANGLED SPEECH REPRESENTATION LEARNING FOR ACOUSTIC MODELLING

Disentangled speech representation learning is based on the assumption that speaker characteristics vary less within an utterance than the linguistic content does, while linguistic content tends to have similar amounts of variation within and across utterances [14]. The FHVAE model [14], which learns to factorise segment-level and sequence-level attributes of sequential data into different latent variables, is applied to disentangle phonetic (linguistic) and speaker information in the speech signal.

The FHVAE's encoder encodes input speech data into segment-level (expected to capture phonetic information) and sequence-level (expected to capture speaker information) latent variables separately, and the FHVAE's decoder reconstructs the original speech based on both the segment- and sequence-level latent variables [14]. Mathematically, let $\boldsymbol{z_1}$ and $\boldsymbol{z_2}$ denote the latent segment variable and the latent sequence variable, respectively. $\boldsymbol{\mu_2}$ is the sequence-dependent prior[8], named as *s-vector*. θ and ϕ denote the parameters of the generation (decoder) and the inference (encoder) models of the FHVAEs, respectively. Let $\mathcal{D} = \{\boldsymbol{X^i}\}_{i=1}^{M}$ denote a speech dataset with M sequences. Each $\boldsymbol{X^i}$ contains N^i speech segments $\{\boldsymbol{x^{(i,n)}}\}_{n=1}^{N^i}$, where $\boldsymbol{x^{(i,n)}}$ contains a number of consecutive frames.

[8] Conceptually analogous to the i-vector in speaker recognition, one vector corresponding to a sequence.

The joint probability for the FHVAE decoder to generate $\boldsymbol{X}$ is formulated as,

$$p_\theta(\boldsymbol{\mu_2}) \prod_{n=1}^{N} p_\theta(\boldsymbol{z_1^n}) p_\theta(\boldsymbol{z_2^n}|\boldsymbol{\mu_2}) p_\theta(\boldsymbol{x^n}|\boldsymbol{z_1^n}, \boldsymbol{z_2^n}). \tag{3.2}$$

In the FHVAE, the exact posterior inference is intractable. The FHVAE introduces an inference model q_ϕ to approximate the intractable true posterior as,

$$q_\phi(\boldsymbol{\mu_2}) \prod_{n=1}^{N} q_\phi(\boldsymbol{z_2^n}|\boldsymbol{x^n}) q_\phi(\boldsymbol{z_1^n}|\boldsymbol{x^n}, \boldsymbol{z_2^n}). \tag{3.3}$$

Details of the formulation of Equations (3.2) and (3.3) are described in Section 3.7.1. The FHVAE model is trained by optimising a discriminative segmental variational lower bound (see Equation (3.10) in Section 3.7.1). To let $\boldsymbol{z_2}$ learn speaker-dependent feature representations, speech utterances of the same speaker in the training data are concatenated into a single sequence before training the FHVAE model. By this means, i, originally defined as the sequence index, becomes equal to the speaker index, and $\boldsymbol{z_1}$ learns a speaker-independent representation.

The use of $\boldsymbol{z_1}$ features in acoustic modelling for oral cancer speech is illustrated in Figure 3.1 (bottom). The GMM-HMM training and the training data are exactly the same as for the fMLLR speaker adaptation method (see Section 3.3.3). The FHVAE model training is implemented using open-source software developed by Hsu et al. [14]. We used FHVAE parameters in [30] in our experiments: The encoder and decoder of the FHVAE are both 2-layer LSTMs with a layer dimension of 256. The dimensions of $\boldsymbol{z_1}$ and $\boldsymbol{z_2}$ are 32. The input features to the FHVAE are fixed-length (10 frames) speech segments. Each frame is represented by a 13-dimensional MFCC with cepstral mean normalisation at the speaker level. During the inference of the $\boldsymbol{z_1}$ features, the FHVAE input segments are shifted by 1 frame, in order to match the length between speech frames and inferred $\boldsymbol{z_1}$.

After FHVAE model training, the $\boldsymbol{z_1}$ features of the WSJ and oral cancer speech are extracted and used as input features for the DNN AM training (see Table 3.3). This system is referred to as *FHVAE*. Compared with the **fMLLR for AM training** system, the only difference in the *FHVAE* system is the input representation to the DNN ($\boldsymbol{z_1}$ versus fMLLR).

3.3.5. PHONEME AND ARTICULATORY FEATURE ANALYSIS

In this section, we describe the error analysis of our trained ASR systems. As a reminder, these analyses have five aims:

- (1) to investigate if the errors made by the ASRs are the same as the known articulation problems in oral cancer speech;
- (2) to find which sounds are poorly/well recognised in the proposed ASR system and to find out which sounds are better recognised with hybrid/E2E architectures
- (3) to compare the errors of the ASR models on healthy and oral cancer speech;
- (4) to provide input to the design of voice commands for ASR systems used by oral cancer speakers.

In our analyses, we will use the phoneme error rate (PER), and the articulatory feature error rate (AFER) as error measures. These metrics are similar to the word error rate (WER), except that they are calculated and interpreted at the level of phonemes and articulatory features (see Section 3.3.5). Confusion matrices of each model will be created and compared with one another to answer our research questions.

Specifically, for our first aim, we are going to look at the worst-performing phonemes and AFs of the *Baseline* system. This analysis assumes that the errors the ASR makes are based on the pronunciation mismatch between oral cancer and WSJ speakers.

For (2), we will investigate which phonemes are consistently misrecognised in the different approach and architecture combinations, and will compare thems in terms of PER and WER. We will investigate whether the different approaches show problems with specific (groups of) phonemes by analysing whether the models have problems capturing particular articulatory feature information by looking at confusion matrices of AFs, or whether these systems' performances are mostly data dependent. We are going to further compare the differences between the best performing Hybrid and E2E techniques. This comparison will allow us to investigate which sounds are better handled by the E2E architectures, and which sounds are better with Hybrid.

For (3), we are going to compare the PER and AFER performances of the E2E and Hybrid Baseline models on the oral cancer and the WSJ tet set. We will denote the WSJ test set experiments as *Hybrid on Healthy* and *E2E on Healthy*. The comparative analysis will allow us to investigate whether the same phonemes are found relatively difficult to the ASR systems.

For (4), we are going to compare the approaches in terms of PER and AFER, and we are interested in which phonemes are recognised well. Phonemes that are recognised well should be preferred in voice commands.

The complete code for the analyses can be found online[9].

PHONEME ERROR RATES AND ARTICULATORY FEATURE ERROR RATE

The PER is calculated as follows. First, the reference (ground truth) sentences and the sentences predicted by the ASR (hypothesis) are converted to phoneme sequences using the CMUdict[10]. The CMUdict contains the ARPABET phonemic transcription of 133,896 English words. Note that we do not take stress into account: Vowels with different stress markers are all treated as the same vowel. Second, the ground-truth phoneme sequence and the hypothesised phoneme sequence are aligned using the Levenshtein distance. We call these alignments Levenshtein alignments. Then, the PER is usually defined as:

$$\text{PER} = \frac{\text{insertion} + \text{substitution} + \text{deletion}}{N}, \tag{3.4}$$

where N is the total number of phonemes in the ground truth phoneme sequence. We also calculate the PER for each individual phoneme f in question as:

[9] https://github.com/karkirowle/relative_phoneme_analysis

[10] In this method, we assume that any errors we observe at the phoneme level are due to the misrecognition of an individual phoneme (leading to a misrecognised word) rather than due to the misrecognition of a word which then would lead to the misrecognition of the phoneme. As we are using a large lexicon for training the ASR (see Section 3.3.1), we think this assumption is reasonable.

$$\text{PER}_\text{f} = \frac{\text{insertion}_\text{f} + \text{substitution}_\text{f} + \text{deletion}_\text{f}}{N_\text{f}}. \tag{3.5}$$

The AFER is calculated similarly to the PER, the main difference being that the aligned phoneme sequences are converted to place of articulation (PoA) and manner of articulation (MoA) feature sequences following Table 3.4 prior to the calculations of the error rates. The AFERs are also reported with respect to each individual articulatory feature, i.e., for the plosives,

$$\text{AFER}_\text{plosives} = \frac{\text{insertion}_\text{plosives} + \text{substitution}_\text{plosives} + \text{deletion}_\text{plosives}}{N_\text{plosives}}. \tag{3.6}$$

We report the mean and standard deviations of PER and AFER over all five test set partitions. In these analyses, we focus on those phonemes that have on average at least 100 occurrences ($N = 100$) in the ground truth, as we believe that 100 occurrences are the bare minimum to make meaningful conclusions. When $N \leq 100$, the results might be influenced too much by data scarcity.

Table 3.4: PoA (columns) and MoA (rows) for each phoneme. **Abbreviations from left to right**: Bilabial, Labiodental, Dental, Alveolar, Postalveolar, Palatal, Velar, Glottal.

PoA / **MoA**	**B**	**LD**	**D**	**A**	**P**	**PAL**	**V**	**G**
Plosives	p,b			t,d			k,g	
Nasal	m			n			ng	
Fricative		f,v	th,dh	s,z		sh,zh		hh
Affricate						jh,ch		
Approximant	w			l	y	r		

CONFUSION MATRICES

Confusion matrices are used in the error analyses to investigate which articulatory feature classes are difficult for the ASRs to capture and which articulatory features are easily confused (modelling error). Using the Levenshtein alignments, we obtain an alignment of the ground truth phoneme sequences and the hypothesised phoneme sequences and create confusion matrices of the phoneme misrecognitions. In our description of the results, we group the phonemes by their AFs.

Since we are interested in the improvement or degradation of AFs in the trained systems compared to the *Baseline*, the Baseline confusion matrix will be separately shown in absolute terms. For the other systems' confusion matrices, the *Baseline* absolute performance will be subtracted.

3.3.6. NOISE ANALYSIS

In this section, we describe our analysis which aims to quantify the influence of noise versus speech severity on the per-recording WER performance.

When quantifying the amount of noise in an audio file, usually the signal-to-noise ratio (SNR) is the figure of interest. Most existing SNR estimation methods are based on measuring the energy content of speech and non-speech regions in a signal. In the case of pathological speech, it has previously been shown that Parkinson's speech and whispered speech can negatively affect the SNR estimation [41]. In other words, it is possible to obtain low SNR estimates in pathological voices even though there is no real background noise present in the recordings .

In order to avoid quantifying noise level by an SNR estimation algorithm that is heavily influenced by the severity of the pathological speech, we wanted to ensure that the correlation between the SNR and severity is low. In order to do that, first, the speech severity of each recording was quantified by an expert listener. To that end, an American English speech language pathologist (SLP) was asked to rate the severity of each recording on a 5-point Likert scale (1: very severe speech, 5: healthy speech) by listening to (at least one) 10 second segment of a recording. (Note that the 10 second segment constraint is based on constraints from an on-going study for which these ratings have been originally collected). The important consequence from the perspective of our analysis is that for some utterances the ratings have higher resolution. By resolution, we mean the step size of MOS during ratings, using one rating only 1-2-3-4-5 is obtainable (step size of 1), using two utterances it is possible to obtain 1-1.5-2-2.5-3-3.5-4-4.5-5 (step size of 0.5). This is because in the case of multiple ratings, we take the mean of the ratings.

Next, for the calculation of the SNR, the gold standard NIST algorithm is used. The NIST SNR is calculated as follows. First, a signal energy histogram is calculated by computing the root mean square (RMS) in dB over a 20 ms analysis window, with a time shift of 10 ms. Typically, this results in a bimodal histogram, one peak (left-most) corresponding to the noise level, and the other peak (right-most) corresponding to the signal level. A raised cosine function is fitted to the noise peak with a direct search algorithm [42], with the objective to minimise the Chi-squared distance. The midpoint of the raised cosine function is labelled as the mean noise power level. The raised cosine curve is then subtracted from the complete RMS histogram to obtain a "noiseless" histogram with a single peak. Then, the peak corresponding to the 95th percentile is defined to be the speech level. Subtracting the noise level from the speech level, the signal to noise ratio is obtained.

Subsequently, Spearman's correlation was calculated between the severity scores and the SNR level ($r = 0.12$, $p \geq 0.5$). The obtained low correlation means that the severity scores and the SNR level are not correlated, therefore the SNR values seem to be independent of the influence of speech severity. This means that our SNR estimates can be reliably used to estimate noise in the recordings.

Finally, to assess the influence of noise on the WER, we did a Pearson's correlation of the per-recording WER (mean across all test partitions) with the SNR for each experiment (SNR-WER r). We perform this analysis for each approach and architecture combination in the paper to see if there are architecture-specific differences in the influence of noise. Furthermore, to assess the influence of speech severity on the WER, we performed a Spearman's correlation of the per-recording WER with the speech severity score (SLP-WER ρ).

Table 3.5: The word error rates (% WER) on the oral cancer speech on the different training-test partitions separately and averaged over all five partitions. **Bold**: best performance among the five systems. For the *Baseline* and *Transformer E2E baseline* systems, both training and test oral cancer speech data are unseen to the system, while for the remaining systems, the oral cancer speech training data is seen to the systems but not the oral cancer speech test data.

System \ Partition index	1		2		3		4		5		Average	
	Training	Test	Training	Test	Training	Test	Training	Test	Training	Test	Training	Test
Hybrid Baseline	78.6	59.7	74.2	75.7	74.0	76.8	74.5	74.9	76.8	65.6	75.6	70.6
Hybrid DNN AM retraining	44.3	55.8	34.7	68.7	40.8	69.5	39.3	71.2	49.3	64.2	41.7	65.9
Hybrid Baseline+OC	53.6	55.8	49.8	74.7	47.5	73.2	47.5	70.9	51.2	62.0	49.9	67.3
fMLLR for AM training	49.0	**52.2**	49.7	69.4	47.4	68.7	46.2	**68.1**	48.9	**55.7**	48.2	**62.8**
FHVAE	50.3	58.0	48.7	73.1	47.0	73.5	46.5	72.6	48.1	65.2	48.1	68.5
E2E baseline	78.6	62.0	74.9	75.5	74.6	76.6	73.6	80.1	76.7	68.3	75.7	72.5
E2E ASR retraining	**23.1**	53.4	**21.8**	**66.0**	**22.3**	**66.4**	**23.8**	71.0	**24.3**	58.0	**23.1**	63.0

3.4. RESULTS AND DISCUSSION

3.4.1. ASR RESULTS

In Section 3.4.1, we first discuss the experimental results of the first five systems listed in Table 3.5, all of which adopt a hybrid DNN-HMM ASR architecture. Next, in Section 3.4.1, we discuss the experimental results of the retraining approach applied to the hybrid versus E2E ASR architectures.

HYBRID ASR RESULTS ON THE TRAINING AND TEST SETS

The word error rates (% WER) on the oral cancer speech data achieved by the *Hybrid Baseline* ASR system, the *Hybrid Baseline+OC* system and the three proposed hybrid systems discussed in Sections 3.3.2, 3.3.3 and 3.3.4 are shown in the top rows in Table 3.5. For each system, the training and test WER results are listed for each of the five training-test data partitions separately (see Table 3.1 for details) and averaged over all partitions. The training WER results are calculated only on the oral cancer training data. **Bold** results indicate the best performance on a particular partition or on the average of all partitions.

Table 3.5 shows that the *Hybrid Baseline* system has the highest training and test WER results of all the systems on all the data partitions (excluding E2E systems). Considering that the *Hybrid Baseline* system achieved a WER of 6.7% on normal speech (see Section 3.3.1), the high WER results for the *Hybrid Baseline* system indicate a severe mismatch between oral cancer speech collected for this study and speech in the WSJ corpus. Although there are several differences between the WSJ and the oral cancer data set (including recording conditions and speaking style (read speech vs. spontaneous speech)). The primary cause of this deterioration is most likely the difference in type of speech, i.e., healthy versus oral cancer speech.

Table 3.5 shows that the *fMLLR* method achieved the best test WER results overall and on four out of the five data partitions (partition 2 is the exception). The hybrid *DNN AM retraining* method achieved an average absolute WER reduction of 34.0% on the training data and of 4.7% on the test data compared to the *Hybrid Baseline* system. The only difference between the AM retraining method and the *Hybrid Baseline* system is the use of a small amount (less than 2 hours, see Table 3.1) of oral cancer speech data during train-

ing in the *DNN AM retraining* system. These results show that such a small amount of speech material already helps to adapt the DNN AM from healthy speech to oral cancer speech and leads to an improvement in recognition performance.

The *fMLLR* system achieved the best performance on the oral cancer test data, achieving an average absolute WER reduction of 7.8% compared to the *Hybrid Baseline* system, and 3.1% compared to the hybrid *DNN AM retraining* system. Not only does the *fMLLR* system outperform the hybrid *DNN AM retraining* system overall on the test data, it also has a better performance on most of the test data partitions (except partition 2). These results suggest that the *fMLLR* approach is better than the DNN AM retraining approach, both in terms of the average WER performance and the per-partition WER performance.

The better performance of the *fMLLR* approach compared to the hybrid *DNN AM retraining* approach is in part due to the merging of the oral cancer speech data with the normal speech data during training, which allows the *fMLLR* model to leverage phonetic information from both healthy speech and oral cancer speech - unlike the *DNN AM retraining* approach which only has access to the oral cancer speech during the retraining phase. A further 4.5% absolute WER reduction on the test data is due to using the fMLLR features (as can be seen when comparing the *fMLLR* system with *Hybrid Baseline+OC*), which allows the model to leverage speaker diversity information.

Interestingly, the hybrid *DNN AM retraining* method achieves the best performance on the training data of all tested systems (excluding E2E systems), but performs worse than the *fMLLR* method on the test data. This finding is likely due to overfitting of the hybrid *DNN AM retraining* method on the small amount of oral cancer training data. At the retraining stage of the *DNN AM retraining* approach, the training data consists of oral cancer speech only. The hybrid DNN AM seems to overfit on the small amount of oral cancer speech training data, which then leads to a less well generalisation to unseen (test) oral cancer speech data. On the other hand, the *fMLLR* method merges the oral cancer speech and normal speech throughout the AM training procedure. In the *fMLLR* approach, during training, the AM is trained to perform well on both the WSJ data and the oral cancer data. This alleviates the overfitting problem, and consequently leads to a better generalisation to unseen oral cancer speech test data compared to the hybrid *DNN AM retraining* method.

The *FHVAE* method achieves better WER performance on the test data than the *Hybrid Baseline* system but worse than the other tested systems. It does achieve the second best WER performance on the training set among all the systems, after the hybrid *DNN AM retraining* method. Notably, the *FHVAE* method performs slightly better than the *Hybrid Baseline+OC* system on the training data, and slightly worse on the test data. The only difference between the *FHVAE* system and *Hybrid Baseline+OC* is the input feature representation to the DNN AM training: the *FHVAE* system uses $\boldsymbol{z_1}$ while the *Hybrid Baseline+OC* uses FBank with pitch features. The comparison between the two systems indicates FHVAE-based disentangled representation learning is effective in alleviating speaker-dependent characteristics in the training data in a limited but consistent manner on all the five partitions. However, it does not generalise well to unseen test data. A possible explanation is the small amount of available oral cancer speech data seen during FHVAE training. In [30], the effectiveness of FHVAE in a low-resource ASR task is shown to be sensitive to the amount of in-domain training data, and was shown to

be very limited when there are only around 2 hours of training data available. To further explore the effect of FHVAE in the oral cancer ASR task, more (unlabelled, as FHVAE is unsupervised) audio recordings from oral cancer speakers should be used, which we leave for future study. However, due to the unlabelled nature of the data, this would be substantially easier to collect in large quantities.

COMPARISON OF THE HYBRID AND E2E ASR ARCHITECTURES IN THE AM RETRAINING APPROACH

The WERs (%) on the oral cancer speech data achieved by the two E2E ASR based systems, i.e., *E2E Baseline* and *E2E ASR retraining*, are shown in bottom rows in Table 3.5. Table 3.5 shows that the *E2E Baseline*'s performance is slightly worse than that of the *Hybrid Baseline* system on both the training and test sets.

Comparison of the two retraining based systems, i.e., *E2E ASR retraining* and the hybrid *DNN AM retraining*, shows that retraining is more effective in the E2E architecture than in the hybrid architecture for the oral cancer ASR task, at least with the current amount of oral cancer retraining data: The absolute WER reduction achieved by retraining is 9.5% for the E2E model, and is 4.7% for the hybrid model. Moreover, the *E2E ASR retraining* system achieves consistently better WER performances across all the partitions than the hybrid *DNN AM retraining* system. The significantly lower training set WERs achieved by the *E2E ASR retraining* indicates stronger modeling capability of the transformer E2E architecture than the hybrid architecture.

The *E2E ASR retraining* system achieves an average test data WER (63.0%) comparable to the best (*fMLLR for AM training*) system which adopts the hybrid ASR architecture (62.8%). Taking all results together, we can conclude that a transformer E2E ASR architecture achieves a WER for oral cancer ASR that approaches but does not outperform the speaker adaptation based hybrid DNN-HMM system.

3.4.2. PHONEME AND ARTICULATORY FEATURE ERROR ANALYSIS

In this section, we present the key results of the error analysis. Each subsection will try to answer one of the five research questions outlined in Section 3.1 and Section 3.3.5. All analyses have been carried out on the five oral cancer speech test set partitions separately and then averaged.

WHAT PHONEMES ARE DIFFICULT FOR THE BASELINE ASR SYSTEMS?

In order to answer this question, we will first look at the phoneme level results, followed by the articulatory level results of the baseline models. Finally, we will compare our results with articulation problems known from the literature.

The phoneme level results are presented in Figure 3.2. The y-axis indicates the PER at the phoneme level. The x-axis shows each of the phonemes in our data set, grouped by manner of articulation. Each line indicates a different system. Shaded regions denote the standard deviation for each model across the 5 folds. As can be seen in Figure 3.2, most phonemes obtain a PER between 40-60%. This indicates that the speech recognition task is challenging. Looking at the blue line (*Hybrid Baseline*), we can identify peaks corresponding to /g/, /aa/, /p/, /th/, /uw/. In the case of the *E2E Baseline*, the most difficult phonemes are /g/, /th/, /uw/, /aa/, /ey/. These are the most difficult phonemes

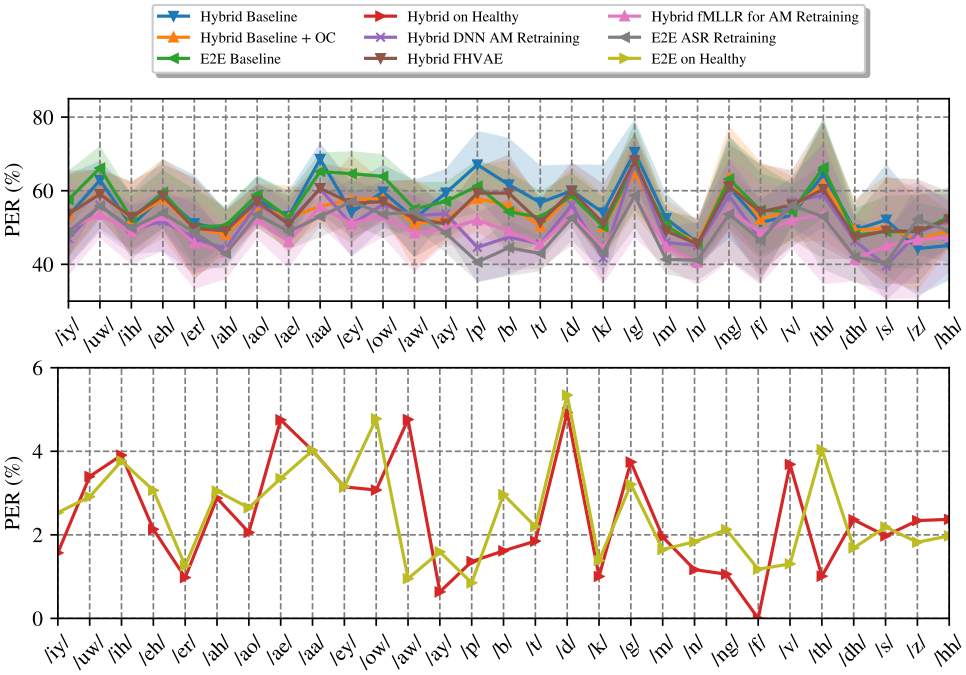


Figure 3.2: Mean PER of each individual phoneme with $n \geq 100$. Shaded regions denote the standard deviation across the 5 folds. Line graph is used for ease of reading. Top panel describes PERs for the oral cancer dataset, while bottom panel describes PERs for the WSJ test set.

for the baseline ASR systems to recognise. We can see that with the exception of /p/ and /ey/, the systems find the same phonemes difficult.

The AF level results are presented in Figure 3.3 and 3.4 (top panels). In the case of the *Hybrid Baseline* MoA, affricates have the highest error, followed by plosives, approximants, nasals, fricatives and then vowels. For PoA, palatal sounds are the worst captured, followed by velars, postalveolars, bilabials, dentals, labiodentals, alveolars, glottals and, finally, vowels. In the case of the *E2E Baseline* MoA, we observe the same order as in the case of *Hybrid Baseline*. For *E2E Baseline* PoA, the palatals are the worst, followed by glottals, labiodentals, dentals, velars, postalveolars, bilabials, alveolars and vowels.

Previous research has already indicated that particularly plosives [4, 5], sibilants [6] and some vowels (/aa/, /ih/, /uw/) [43, 44] are impacted by oral cancer. We can see that plosives have the second worst AFER, with two plosives (/g/ and /p/) having a PER of over 60%. As for sibilants (in our analysis: (post)alveolar fricatives), we observe that /s/ and /z/ are both comparatively well captured by the baseline ASR systems, showing that our systems did not have relatively more difficulty capturing sibilant information compared to other groups of phonemes. Finally, we can see that vowels are relatively well captured, with the exception of /aa/ and /uw/, which is consistent with the literature. The difficulty in recognising words with /ih/ as indicated by [43] is not observed.

Overall, we see that those sounds that are known to cause articulatory problems after

surgery for oral cancer speech, are also hard to recognise for the baseline ASR systems we tested. This is particularly the case for plosives and two vowels /aa/ and /uw/. In deviance to the literature, our systems did not have particular problems with sibilants. The reason for this difference is unclear: it might be that ASRs are more robust to variations in sibilant realisation. It would be interesting to confirm this with lisping speakers, where only sibilants are impacted. Interestingly, there were no sounds or articulatory features that were relatively hard for our ASR systems that were as yet unknown in the literature.

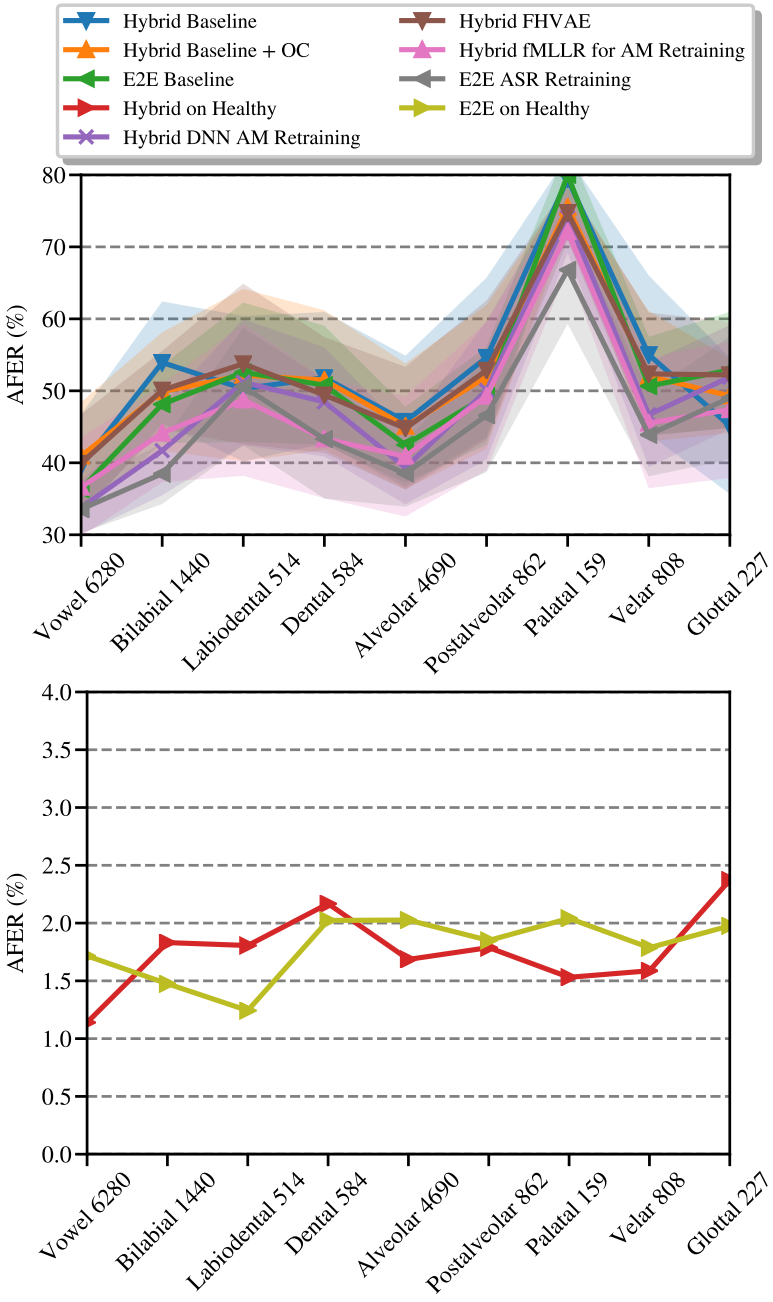


Figure 3.3: **Top**: Comparison of AFER for PoA on the oral cancer test set. **Bottom**: Comparison of AFER for PoA on the WSJ test set. Mean N (phonemes in test set) rounded to three significant figures are in parentheses.

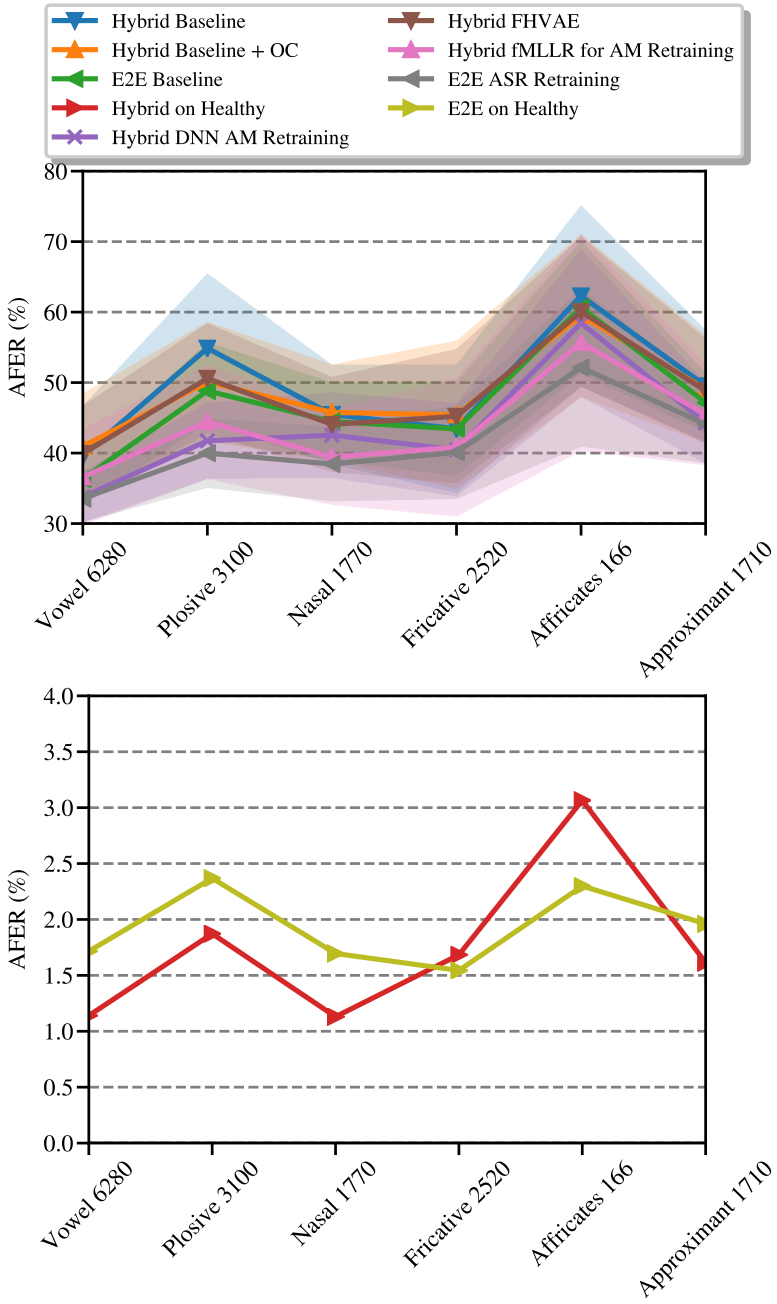


Figure 3.4: **Top**: Comparison of AFER for MoA on the oral cancer test set. **Bottom**: Comparison of AFER for MoA on the WSJ test set. Mean N (phonemes in test set) rounded to three significant figures are in parentheses.

Figure 3.5: Relative confusion matrices on PoA (top) and on MoA (bottom). Green diagonals and red off-diagonals mean better performance, while red diagonals and green off-diagonals mean worse performance. Green background denotes the absolute performances, while white background denotes relative performances.

HOW WELL/POORLY PHONEMES ARE RECOGNISED IN THE PROPOSED ASR SYSTEMS?

In order to investigate what techniques lead to a good recognition performance of oral cancer speech and what needs further investigation, we investigate which phonemes are improved and which ones are still misrecognised by analysing the produced error rates. Both for the phoneme and articulatory feature analysis, we additionally list the phonemes which seemed to work better with E2E architecture, and those which seemed to work better with Hybrid architecture.

As Figure 3.2 shows, overall, the individual PER lies between 40% and 60%. A comparison of the different models shows that all the approaches generally improve the individual PERs compared to the baseline models (the blue line for the hybrid Baseline model and the green line for the E2E Baseline model), with a few exceptions, most notably the /hh/, /z/, /f/, /ey/ where particularly the *Hybrid Baseline*+OC orange line) and *FHVAE* (red line) models perform worse than the *Hybrid Baseline* model. In the case of the *E2E Baseline*, Hybrid Baseline+OC and FHVAE trained models perform worse on /b/. The hybrid systems outperform the E2E model in the case of /iy/, /uw/, /ih/, /eh/, /er/, /ao/, /ae/, /ey/, /ow/, /aw/ (therefore with most vowels), /k/, /g/, /v/, /dh/, /z/, and /hh/. The *E2E ASR retraining* system is better with /ah/, /aa/, /ay/, /p/, /b/, /t/, /d/ (therefore with most plosives), /m/, /n/ (all of the nasals), /g/, /f/, /th/, and /s/.

To further investigate whether certain (groups of) phonemes are consistently misrecognised, we investigate whether there is particular articulatory feature information that the models do not capture well. The extent to which the models can capture articulatory feature information is visualised in Figure 3.4 and 3.3, the x-axis showing the different MoA/PoA and the number of phonemes (n) in each class, the y-axis showing the AFER. For PoA, palatal, postalveolar and velar sounds seem to be the most challenging, while for MoA these are affricates and approximants. Although all models in general improved the uptake of articulatory feature information (glottals being the exception), this was particularly the case for the Hybrid DNN AM retraining/E2E ASR retraining models for bilabial and plosive information. We observe that E2E better captures bilabial, alveolar, postalveolar, palatal, and velar information, while vowels, labiodentals, dentals and glottals are better captured by the hybrid models. For MoA (Figure 3.4), the E2E model better captures plosive, nasal, fricative, affricate and approximant information, while vowel information is slightly better captured by the hybrid models, which actually has a larger impact on the overall performance (this can be observed by looking at the number of phonemes in each category, which is in parentheses).

We were interested if the difference between the AFER performances (i.e., vowels vs. affricates) was due to data scarcity in the phoneme classes to which the AFs were underlying. To investigate this, we performed a post-hoc Pearson's correlation analysis between the number of phonemes (of the AF class) (n) as the independent variable, and the PER performance as the dependent variable. The analysis found relatively strong effect sizes (*Hybrid DNN AM retraining*: 0.51, *fMLLR*: 0.55, *E2E ASR retraining*: 0.52, $p \leq 0.01$). Along with the fact that nearly all phonemes improve with our three approaches and for both architectures, we can conclude that the bottleneck of the performance seems to be mostly data-dependent. This means that it is important to collect corpora for oral cancer ASR in a phonetically balanced way, in order to have enough data to build good sound representations of each phoneme, including the rarer phonemes, such as glottals

and palatals.

The confusion matrices in Figure 3.5 enable further interpretation of these results. To better visualise the improvements, we have used relative confusion matrices for the proposed systems. In the case of relative confusion matrices, a green diagonal (more correct class) and a red off-diagonal (fewer incorrect classes) means improved classification. Also, note that for the insertion and deletion errors, a white line (meaning no errors) would be ideal for the absolute case, and a red or white line for the relative case (decreased errors or no change).

As a general remark, we can see that the majority of improvements in the *fMLLR* and *Hybrid DNN AM retraining* come from the reduction of deletion errors (red vertical line on the left side of the plot). For MoA, an additional part of this improvement comes from a reduction in substitutions of plosive sounds with fricative sounds compared to the *Hybrid Baseline*. Regarding PoA, we can see that (mainly) alveolar sounds and vowels were substituted with glottal sounds in the *Hybrid Baseline* model (light green vertical line in the middle), which is alleviated in the proposed approaches (red vertical lines in the middle of the plots). In the case of the *E2E ASR retraining model* we observe that fewer sounds are classified as glottals, which makes the performance of the model worse on glottals overall compared to the *Hybrid DNN AM retraining*. Furthermore, a lot of phonemes are misclassified as dentals - it can be observed (vertical green lines) that the *E2E ASR retraining* model seems to make dentals as the "fallback" articulatory feature category.

We can summarise the findings as follows: (a) Plosive sounds are impacted in oral cancer speech, but speaker-adaptive training (*fMLLR* and *FHVAE*) and even a relatively small amount of training data (2 hours; all proposed approaches) seem to alleviate these problems with the recognition of plosives. (b) Performance seems to be heavily data dependent, in general the number of phonemes is a good predictor of performance. (c) The "recognition" of /z/ and /hh/ is not improved over *Hybrid Baseline*, however this is partially explained by (b) as these two phoneme classes have relatively small amounts of training data (/z/ = 373 occurrences, /hh/ = 227 occurrences). This means that data augmentation techniques could be useful to alleviate the data scarcity problem. Overall, PER improvements brought by the proposed approaches compared to the baseline systems can be attributed to a general improvement in recognition performance across all phonemes. (d) In terms of manner of articulation, hybrid is only better compared to E2E on vowels - however, vowels have a large contribution to overall performance. As for place of articulation, vowels, labiodentals, dentals and glottals are better captured by hybrid models, while E2E better capture bilabial, alveolar, postalveolar, palatal and velar information.

Thus, in order to improve ASR for oral cancer speech, we conclude: (a) retraining approaches with even a small amount of extra training data can lead to substantial improvements for the AM; (b) Data augmentation techniques should be investigated for oral cancer ASR.

DO MISRECOGNITIONS OF ORAL CANCER PHONEMES COINCIDE WITH MISRECOGNITIONS OF HEALTHY PHONEMES?

In this section, we would like to answer the question of whether the phoneme errors of the different approaches and architectures on oral cancer speech coincide with their er-

rors on typical, healthy speech. In order to do that, we compare the PERs and the AFERs of the two baseline architectures (Hybrid and E2E Baseline) on both the oral cancer test set and the WSJ test set. (Note that this analysis is only carried out using the Baseline models as these are the only models that are only trained on healthy speech.)

The PERs on the oral cancer speech can be seen in the top panel of Figure 3.2, while the PERs of the healthy speech can be seen in the bottom panel. We consider a phoneme relatively badly recognised in the case of oral cancer speech when the PER is over 60%. In the case of healthy speech, we set a threshold of 4%.

In the case of the hybrid architecture tested on healthy speech (*Hybrid on Healthy*) the phonemes /ae/, /aa/, /aw/ and /d/ are above the 4% threshold. In the case of the E2E architecture tested on healthy speech (*E2E on Healthy*), /aa/, /ow/, /d/ and /th/ are above the 4% threshold. For the hybrid architecture tested on oral cancer speech (*Hybrid Baseline*), the phonemes /uw/, /aa/, /p/, /b/, /g/, /ng/, /th/ are above the 60% threshold. For the E2E architecture tested on oral cancer speech (*E2E Baseline*), /uw/, /aa/, /ey/, /ow/, /p/, /g/, /ng/, /th/ are relatively badly recognised.

We can observe the following from these results. (1) The phonemes /aa/ and /d/ are relatively difficult for all architectures, independent of the type of speech used. (2) The phonemes /uw/, /p/, /g/, /ng/ are relatively more difficult in the case of oral cancer speech than in healthy speech.

This last finding (2) is partially consistent with the literature results discussed in Section 3.4.2, with the exception of /ng/. The /uw, p, and g/ sounds probably have a different pronunciation in oral cancer speech compared to healthy speech, leading to a worse recognition of these sounds by the Baseline models which have not been trained on oral cancer speech.

WHAT VOICE COMMANDS SHOULD BE USED WITH ORAL CANCER ASR?

When developing speech-driven systems or oral cancer speakers, it is preferable to base these on either the *Hybrid fMLLR* or *Hybrid DNN AM retraining* approaches as these are the two best systems. The results in the previous two subsections show that the phonemes that are best recognised by the DNN AM retraining are /s, k, ah, p, n/, while for *fMLLR retraining* these are /n, dh, ah, k, m/. So depending on which approach is used, we recommend selecting words containing these phonemes for the voice commands. Note that even though plosives are affected in oral cancer speech, our ASR results do not indicate that plosives should be excluded when designing voice commands.

Table 3.6: Word error rates (%) and signal to noise ratios (SNR in dB) of the recordings in the dataset. Significance levels: * ($p < 0.5$), ** ($p < 0.01$), *** ($p < 0.001$)

Recording id	Baseline	Baseline + OC	DNN AM Retraining	fMLLR	FHVAE	E2E Baseline	E2E ASR retraining	SNR	SLP score
001	76.76	68.94	60.61	56.32	68.18	79.0	57.0	41.0	3.5
003	64.68	62.74	54.0	61.66	63.72	63.1	53.35	55.0	4.5
010	89.67	85.33	86.06	84.55	86.17	100.2	91.6	63.75	2.3
018	47.63	43.31	44.75	38.27	45.83	48.2	41.87	42.25	5.0
021	70.86	68.7	69.07	56.18	69.02	73.3	59.8	42.25	3.88
023	85.8	85.2	78.46	81.2	82.06	85.7	77.6	41.75	2.3
024	88.56	86.7	80.71	81.17	84.59	87.6	76.5	33.25	2.6
030	57.41	49.38	52.47	42.59	53.09	70.4	51.23	29.25	4.5
033	80.39	81.88	68.58	75.37	83.55	83.3	70.75	25.25	4.4
034	62.18	57.71	60.14	55.42	61.18	64.57	56.23	67.0	4.9
SNR-WER r	-0.04	-0.07	0.08	0.07	-0.05	-0.04	0.11	–	–
SLP-WER ρ	-0.93 ***	-0.91 ***	-0.91 ***	-0.87 **	-0.87 **	-0.93 ***	-0.91 ***	–	–

3.4.3. HOW DOES NOISE IN THE DATASET IMPACT THE RESULTS OF THE ASR SYSTEMS?

Table 3.6 shows the influence of noise and speech severity on the WER. Each row corresponds to one audio recording with the corresponding WER rates on the different ASR models. From the low SNR-WER r correlation results, we can see that the impact of noise is generally low on the audio. The highest correlation between the SNR and the WER is for *E2E ASR retraining*, none of the SNR-WER correlations are significant. We can thus conclude that noise does not seem to have an influence on the WER results.

On the other hand, in all experimental conditions the speech severity seemed to be highly and significantly correlated with the WER results. The highest correlation is in the case of the *E2E Baseline*, followed by *Hybrid Baseline+OC*, *Hybrid DNN AM retraining*, *E2E ASR retraining*, and finally the *FHVAE* and the *fMLLR* methods. We can thus conclude that speech severity always has an influence on WER with the largest influence when there is no oral cancer data used for training, and the least influence when speaker-adaptive training is used.

Nevertheless, our subjective impression is that some recordings have quite challenging acoustic conditions for which speech enhancement techniques might be useful. We leave this for future research: for instance, one approach could be for speakers who have multiple recordings (such as id008 and id011) to use a VoiceFilter-based speech enhancement [45]. In that enhancement technique, an auxiliary recording is used to separate channel information pertaining to the speaker and background noise. Because there are many non-stationary noise sources in these audios, the VoiceFilter approach would probably be more beneficial than a spectral subtraction based approach, which is known to remove only stationary noise.

To summarise, we can conclude that speech severity impacts the WER performance to a great extent, and the impact of noise on the WER performance is substantially less.

3.4.4. FUTURE WORK ON THE ROLE OF DATA AUGMENTATION

We hypothesise that some data augmentation techniques (such as pitch shift) would not work in the case of oral cancer speech, as the original speech is often already distorted beyond human comprehensibility. Existing literature for similar speech pathologies propose predominantly specific, custom techniques, i.e., the current state-of-the-art dysarthric ASR system uses speed perturbation [18], other techniques propose voice conversion [46, 21]. In the work of [21], it is also stated that data augmentation approaches seem to work better for high intelligibility pathological speakers. Therefore, we believe analysis of data augmentation techniques warrant a separate study, where effects such as the type of data augmentation, amount of data, and severity of speech can be separated in a controlled way.

3.5. CONCLUSION

In this paper, we presented a new dataset of American English oral cancer speech collected from YouTube. We investigated and compared two different DNN architectures on the task of oral cancer ASR with three different approaches: a *DNN AM retraining* (Hybrid, End-to-End) approach, an *fMLLR for AM training* approach, and an *FHVAE* ap-

proach. The *fMLLR* approach performed the best overall and achieved a WER of 62.8% on the oral cancer speech test set, which is a 7.8% absolute improvement over the *Hybrid Baseline*. Detailed error analyses on the recognition results of these approaches and architectures showed that (1) plosives and some vowels are challenging to recognise for the *Baseline* systems trained without oral cancer data, which is consistent with the literature on oral cancer speech which indicates that particularly plosives and some vowels are impacted by the removal of (parts of) the tongue due to oral cancer speech treatment. In contrast to the oral cancer literature, our models do not show the known problems with sibilants. In other words, we find that ASRs even without seeing oral cancer speech perform relatively well on sibilants of oral cancer speech. (2) The proposed approaches successfully alleviate the problems with the recognition of plosives and vowels. Furthermore, the proposed approaches and architectures do not show problems with particular phonemes, but rather their performance depends on the amount of training data for a given phoneme. Future research should therefore be directed towards data augmentation of particularly those phonemes with less training material, and speech enhancement techniques. (3) We find that it is mainly /uw/, /p/, /g/, /ng/ that are relatively difficult to recognise in the case of oral cancer speech, but not in the case of healthy speech (this analysis was only carried out on the Baseline systems). (4) For the development of voice command systems for oral cancer speakers, we propose to select words that include phonemes /s/, /k/, /ah/, /p/, /n/ for a system based on *Hybrid DNN AM retraining*, and /n/, /dh/, /ah/, /k/, /m/ for a system based on *fMLLR*. (5) A final analysis showed that channel noise in the recordings does not have an impact on the recognition performance of the models, rather the poor performance on the oral cancer speech is caused by the severity of the speech pathology.

3.6. ACKNOWLEDGEMENTS

We would like to thank Noa Hannah (University of Illinois at Urbana-Champaign) for providing the severity ratings. This project has received funding from the European Union's Horizon 2020 research and innovation programme under Marie Skłodowska-Curie grant agreement No 766287 (TAPAS). The Department of Head and Neck Oncology and surgery of the Netherlands Cancer Institute receives a research grant from Atos Medical (Hörby, Sweden), which contributes to the existing infrastructure for quality of life research.

3.7. SUPPLEMENTARY MATERIAL

3.7.1. DETAILS OF THE FHVAE MODEL

We follow the terminology used in [14] to describe the details of the FHVAE model. Let $\mathcal{D} = \{\boldsymbol{X}^{\boldsymbol{i}}\}_{i=1}^{M}$ denote a speech dataset with M sequences. The i-th sequence $\boldsymbol{X}^{\boldsymbol{i}}$ contains N^i speech segments $\{\boldsymbol{x}^{(\boldsymbol{i},\boldsymbol{n})}\}_{n=1}^{N^i}$, where $\boldsymbol{x}^{(\boldsymbol{i},\boldsymbol{n})}$ is a segment of a fixed number of frames. The FHVAE model formulates the generation process of a sequence $\boldsymbol{X}$ as[11] [14],

1. A vector $\boldsymbol{\mu_2}$ is drawn from a prior distribution $p_\theta(\boldsymbol{\mu_2}) = \mathcal{N}(\boldsymbol{0}, \sigma^2_{\boldsymbol{\mu_2}}\boldsymbol{I})$;
2. Latent segment variables $\boldsymbol{z_1^n}$ and latent sequence variables $\boldsymbol{z_2^n}$ are drawn from $p_\theta(\boldsymbol{z_1^n}) = \mathcal{N}(\boldsymbol{0}, \sigma^2_{\boldsymbol{z_1}}\boldsymbol{I})$ and $p_\theta(\boldsymbol{z_2^n}|\boldsymbol{\mu_2}) = \mathcal{N}(\boldsymbol{\mu_2}, \sigma^2_{\boldsymbol{z_2}}\boldsymbol{I})$;
3. Speech segment $\boldsymbol{x^n}$ is drawn from

$$p_\theta(\boldsymbol{x^n}|\boldsymbol{z_1^n}, \boldsymbol{z_2^n}) = \mathcal{N}(f_{\boldsymbol{\mu_x}}(\boldsymbol{z_1^n}, \boldsymbol{z_2^n}), diag(f_{\boldsymbol{\sigma_x^2}}(\boldsymbol{z_1^n}, \boldsymbol{z_2^n})). \tag{3.7}$$

Here $\mathcal{N}$ denotes the standard normal distribution, $f_{\boldsymbol{\mu_x}}(\cdot,\cdot)$ and $f_{\boldsymbol{\sigma_x^2}}(\cdot,\cdot)$ are parameterized by two DNNs. Based on Equation (3.7), the joint probability for generating $\boldsymbol{X}$ is formulated as (same as Equation (3.2)),

$$p_\theta(\boldsymbol{\mu_2}) \prod_{n=1}^{N} p_\theta(\boldsymbol{z_1^n}) p_\theta(\boldsymbol{z_2^n}|\boldsymbol{\mu_2}) p_\theta(\boldsymbol{x^n}|\boldsymbol{z_1^n}, \boldsymbol{z_2^n}). \tag{3.8}$$

The FHVAE introduces an inference model to approximate the true posterior as follows (same as Equation (3.3)),

$$p_\phi(\boldsymbol{\mu_2}) \prod_{n=1}^{N} p_\phi(\boldsymbol{z_2^n}|\boldsymbol{x^n}) p_\phi(\boldsymbol{z_1^n}|\boldsymbol{x^n}, \boldsymbol{z_2^n}). \tag{3.9}$$

Here $p_\phi(\boldsymbol{\mu_2})$, $p_\phi(\boldsymbol{z_2^n}|\boldsymbol{x^n})$ and $p_\phi(\boldsymbol{z_1^n}|\boldsymbol{x^n}, \boldsymbol{z_2^n})$ are all diagonal Gaussian distributions. The mean and variance values of $p_\phi(\boldsymbol{z_2^n}|\boldsymbol{x^n})$ and $p_\phi(\boldsymbol{z_1^n}|\boldsymbol{x^n}, \boldsymbol{z_2^n})$ are parameterized by DNNs.

The FHVAE is trained to optimise the *discriminative segmental variational lower bound* $\mathcal{L}(\theta, \phi; \boldsymbol{x}^{(\boldsymbol{i},\boldsymbol{n})})$ [14], which is defined as,

$$\begin{aligned}
&\mathbb{E}_{q_\phi(\boldsymbol{z_1}^{(\boldsymbol{i},\boldsymbol{n})}, \boldsymbol{z_2}^{(\boldsymbol{i},\boldsymbol{n})}|\boldsymbol{x}^{(\boldsymbol{i},\boldsymbol{n})})}[\log p_\theta(\boldsymbol{x}^{(\boldsymbol{i},\boldsymbol{n})}|\boldsymbol{z_1}^{(\boldsymbol{i},\boldsymbol{n})}, \boldsymbol{z_2}^{(\boldsymbol{i},\boldsymbol{n})})] \\
&- \mathbb{E}_{q_\phi(\boldsymbol{z_2}^{(\boldsymbol{i},\boldsymbol{n})}|\boldsymbol{x}^{(\boldsymbol{i},\boldsymbol{n})})}[\mathrm{KL}(q_\phi(\boldsymbol{z_1}^{(\boldsymbol{i},\boldsymbol{n})}|\boldsymbol{x}^{(\boldsymbol{i},\boldsymbol{n})}, \boldsymbol{z_2}^{(\boldsymbol{i},\boldsymbol{n})})||p_\theta(\boldsymbol{z_1}^{(\boldsymbol{i},\boldsymbol{n})}))] \\
&- \mathrm{KL}(q_\phi(\boldsymbol{z_2}^{(\boldsymbol{i},\boldsymbol{n})}|\boldsymbol{x}^{(\boldsymbol{i},\boldsymbol{n})})||p_\theta(\boldsymbol{z_2}^{(\boldsymbol{i},\boldsymbol{n})}|\tilde{\boldsymbol{\mu}}_{\boldsymbol{2}}^{\boldsymbol{i}})) \\
&+ \frac{1}{N^i}\log p_\theta(\tilde{\boldsymbol{\mu}}_{\boldsymbol{2}}^{\boldsymbol{i}}) + \alpha \log p(i|\boldsymbol{z_2}^{(\boldsymbol{i},\boldsymbol{n})}),
\end{aligned} \tag{3.10}$$

where $\tilde{\boldsymbol{\mu}}_{\boldsymbol{2}}^{\boldsymbol{i}}$ denotes the posterior mean of $\boldsymbol{\mu_2}$ for the i-th sequence and α denotes the discriminative weight. The discriminative objective $\log p(i|\boldsymbol{z_2}^{(\boldsymbol{i},\boldsymbol{n})})$ is formulated as,

$$\log p(i|\boldsymbol{z_2}^{(\boldsymbol{i},\boldsymbol{n})}) := \log p_\theta(\boldsymbol{z_2}^{(\boldsymbol{i},\boldsymbol{n})}|\tilde{\boldsymbol{\mu}}_{\boldsymbol{2}}^{\boldsymbol{i}}) - \log \sum_{j=1}^{M} p_\theta(\boldsymbol{z_2}^{(\boldsymbol{j},\boldsymbol{n})}|\tilde{\boldsymbol{\mu}}_{\boldsymbol{2}}^{\boldsymbol{j}}). \tag{3.11}$$

After FHVAE training, $\boldsymbol{z_1}$ representation is extracted as the desired speaker-invariant representation of speech.

[11] For simplicity, the superscript i in $\boldsymbol{X}^{\boldsymbol{i}}$ and subsequent equations is omitted. This does not cause confusion.

BIBLIOGRAPHY

[1] K. D. Shield, J. Ferlay, A. Jemal, R. Sankaranarayanan, A. K. Chaturvedi, F. Bray, and I. Soerjomataram, "The global incidence of lip, oral cavity, and pharyngeal cancers by subsite in 2012," *CA: A Cancer Journal for Clinicians*, vol. 67, no. 1, pp. 51–64, 2017.

[2] The Oral Cancer Foundation, "Oral cancer facts," Feb 2019. [Online]. Available: https://oralcancerfoundation.org/facts/

[3] E. C. Ward and C. J. van As-Brooks, *Head and Neck Cancer: treatment, rehabilitation, and outcomes. Chapter 5: Speech and Swallowing Following Oral, Oropharyngeal, and Nasopharyngeal Cancer.* Plural Publishing, 2014.

[4] T. Bressmann, H. Jacobs, J. Quintero, and J. C. Irish, "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*, vol. 33, no. 4, p. 204, 2009.

[5] T. Bressmann, R. Sader, T. L. Whitehill, and N. Samman, "Consonant Intelligibility and Tongue Motility in Patients with Partial Glossectomy," *Journal of Oral and Maxillofacial Surgery*, vol. 62, no. 3, pp. 298–303, 2004.

[6] J.-P. Laaksonen, J. Rieger, J. Harris, and H. Seikaly, "A Longitudinal Acoustic Study of the Effects of the Radial Forearm Free Flap reconstruction on Sibilants Produced by Tongue Cancer Patients," *Clinical Linguistics & Phonetics*, vol. 25, no. 4, pp. 253–264, 2011.

[7] B. M. Halpern, R. van Son, M. van den Brekel, and O. Scharenborg, "Detecting and Analysing Spontaneous Oral Cancer Speech in the Wild," in *Proc. Interspeech 2020*, 2020, pp. 4826–4830. [Online]. Available: http://dx.doi.org/10.21437/Interspeech.2020-1598

[8] J. B. Epstein, S. Emerton, D. A. Kolbinson, N. D. Le, N. Phillips, P. Stevenson-Moore, and D. Osoba, "Quality of life and oral function following radiotherapy for head and neck cancer." *Head Neck*, 1999.

[9] J. A. Logemann, B. R. Pauloski, A. W. Rademaker, and L. A. Colangelo, "Speech and Swallowing Rehabilitation for Head and Neck Cancer Patients," *Oncology*, vol. 11, no. 5, 1997.

[10] K. Kappert, M. van Alphen, L. Smeele, A. Balm, and F. van der Heijden, "Quantification of Tongue Mobility Impairment Using Optical Tracking in Patients After Receiving Primary Surgery or Chemoradiation," *PloS one*, vol. 14, no. 8, 2019.

[11] M. Windrich, A. Maier, R. Kohler, E. Nöth, E. Nkenke, U. Eysholdt, and M. Schuster, "Automatic Quantification of Speech Intelligibility of Adults with Oral Squamous Cell Carcinoma," *Folia Phoniatrica et Logopaedica*, vol. 60, no. 3, pp. 151–6, 04 2008.

[12] H. Xu, V. H. Do, X. Xiao, and E. S. Chng, "A Comparative Study of BNF and DNN Multilingual Training on Cross-Lingual Low-Resource Speech Recognition," in *Interspeech*, 2015, pp. 2132–2136.

[13] M. Heck, S. Sakti, and S. Nakamura, "Feature Optimized DPGMM Clustering For Unsupervised Subword Modeling: A contribution to zerospeech 2017," in *Proc. ASRU*. IEEE, 2017, pp. 740–746.

[14] W. Hsu, Y. Zhang, and J. R. Glass, "Unsupervised Learning of Disentangled and Interpretable Representations from Sequential Data," in *Proc. NIPS*, 2017, pp. 1878–1889.

[15] H. Christensen, M. B. Aniol, P. Bell, P. D. Green, T. Hain, S. King, and P. Swietojanski, "Combining in-domain and out-of-domain speech data for automatic recognition of disordered speech," in *Interspeech*, 2013, pp. 3642–3645.

[16] Y. Liu, T. Lee, P. C. Ching, T. K. T. Law, and K. Y. S. Lee, "Acoustic Assessment of Disordered Voice with Continuous Speech Based on Utterance-Level ASR Posterior Features," in *Interspeech*, 2017, pp. 2680–2684.

[17] E. Yilmaz, M. Ganzeboom, C. Cucchiarini, and H. Strik, "Multi-stage DNN Training for Automatic Recognition of Dysarthric Speech," in *Interspeech*, 2017, pp. 2685–2689.

[18] E. Hermann and M. M. Doss, "Dysarthric Speech Recognition with Lattice-Free MMI," in *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2020, pp. 6109–6113.

[19] Y. Qin, T. Lee, S. Feng, and A. P. Kong, "Automatic Speech Assessment for People with Aphasia Using TDNN-BLSTM with Multi-Task Learning," in *Interspeech*, 2018, pp. 3418–3422.

[20] S. Karita, X. Wang, S. Watanabe, T. Yoshimura, W. Zhang, N. Chen, T. Hayashi, T. Hori, H. Inaguma, Z. Jiang, M. Someki, N. E. Y. Soplin, and R. Yamamoto, "A Comparative Study on Transformer vs RNN in Speech Applications," in *Proc. ASRU*, 2019, pp. 449–456.

[21] J. Harvill, D. Issa, M. Hasegawa-Johnson, and C. Yoo, "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)*. IEEE, 2021, pp. 6428–6432.

[22] T. Anastasakos, J. McDonough, R. Schwartz, and J. Makhoul, "A Compact Model for Speaker-Adaptive Training," in *Proc. ICSLP*, vol. 2, 1996, pp. 1137–1140.

[23] V. Gupta, P. Kenny, P. Ouellet, and T. Stafylakis, "I-vector-based Speaker Adaptation of Deep Neural Networks For French Broadcast Audio Transcription," in *ICASSP*. IEEE, 2014, pp. 6334–6338.

[24] T. Anastasakos, J. McDonough, and J. Makhoul, "Speaker Adaptive Training: A Maximum Likelihood Approach to Speaker Normalization," in *ICASSP*, vol. 2. IEEE, 1997, pp. 1043–1046.

[25] Y. Miao, H. Zhang, and F. Metze, "Speaker Adaptive Training of Deep Neural Network Acoustic Models using i-vectors," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 23, no. 11, pp. 1938–1949, 2015.

[26] X. Cui, V. Goel, and G. Saon, "Embedding-Based Speaker Adaptive Training of Deep Neural Networks," in *Proc. Interspeech 2017*, 2017, pp. 122–126.

[27] M. J. Gales, "Maximum Likelihood Linear Transformations for HMM-based Speech Recognition," *Computer Speech & Language*, vol. 12, no. 2, pp. 75–98, 1998.

[28] S. Hahm, D. Heitzman, and J. Wang, "Recognizing Dysarthric Speech due to Amyotrophic Lateral Sclerosis with Across-speaker Articulatory Normalization," in *Proceedings of SLPAT 2015: 6th Workshop on Speech and Language Processing for Assistive Technologies*, 2015, pp. 47–54.

[29] C. Bhat, B. Vachhani, and S. K. Kopparapu, "Recognition of Dysarthric Speech Using Voice Parameters for Speaker Adaptation and Multi-Taper Spectral Estimation," in *Interspeech*, 2016, pp. 228–232.

[30] S. Feng and T. Lee, "Improving Unsupervised Subword Modeling via Disentangled Speech Representation Learning and Transformation," in *Proc. INTERSPEECH*, 2019, pp. 281–285.

[31] S. Feng, T. Lee, and Z. Peng, "Combining Adversarial Training and Disentangled Speech Representation for Robust Zero-Resource Subword Modeling," in *Proc. INTERSPEECH*, 2019, pp. 1093–1097.

[32] W. Hsu and J. R. Glass, "Extracting Domain Invariant Features by Unsupervised Learning for Robust Automatic Speech Recognition," in *Proc. ICASSP*. IEEE, 2018, pp. 5614–5618.

[33] X. Cui and A. Alwan, "Noise robust speech recognition using feature compensation based on polynomial regression of utterance SNR," *IEEE Transactions on Speech and Audio Processing*, vol. 13, no. 6, pp. 1161–1172, 2005.

[34] D. B. Paul and J. M. Baker, "The Design for the Wall Street Journal-based CSR Corpus," in *Proceedings of the workshop on Speech and Natural Language*. Association for Computational Linguistics, 1992, pp. 357–362.

[35] P. Ghahremani, B. BabaAli, D. Povey, K. Riedhammer, J. Trmal, and S. Khudanpur, "A Pitch Extraction Algorithm Tuned for Automatic Speech Recognition," in *ICASSP*. IEEE, 2014, pp. 2494–2498.

[36] D. Povey, A. Ghoshal, G. Boulianne, L. Burget, O. Glembek, N. Goel, M. Hannemann, P. Motlicek, Y. Qian, P. Schwarz *et al.*, "The Kaldi Speech Recognition Toolkit," in *Proc. ASRU*, 2011, pp. 1–4.

[37] K. Veselỳ, A. Ghoshal, L. Burget, and D. Povey, "Sequence-Discriminative Training of Deep Neural Networks," in *Interspeech*, 2013, pp. 2345–2349.

[38] D. Povey, V. Peddinti, D. Galvez, P. Ghahremani, V. Manohar, X. Na, Y. Wang, and S. Khudanpur, "Purely Sequence-Trained Neural Networks for ASR Based on Lattice-Free MMI," in *Proc. INTERSPEECH*, 2016, pp. 2751–2755.

[39] N. Zeghidour, Q. Xu, V. Liptchinsky, N. Usunier, G. Synnaeve, and R. Collobert, "Fully convolutional speech recognition," *arXiv preprint arXiv:1812.06864*, 2018.

[40] S. Kim, T. Hori, and S. Watanabe, "Joint CTC-attention based end-to-end speech recognition using multi-task learning," in *Proc. ICASSP*. IEEE, 2017, pp. 4835–4839.

[41] A. H. Poorjam, M. A. Little, J. R. Jensen, and M. G. Christensen, "A supervised approach to global signal-to-noise ratio estimation for whispered and pathological voices," in *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2018, pp. 296–300.

[42] R. Hooke and T. A. Jeeves, ""Direct Search"Solution of Numerical and Statistical Problems," *Journal of the ACM (JACM)*, vol. 8, no. 2, pp. 212–229, 1961.

[43] J. Takatsu, N. Hanai, H. Suzuki, M. Yoshida, Y. Tanaka, S. Tanaka, Y. Hasegawa, and M. Yamamoto, "Phonologic and acoustic analysis of speech following glossectomy and the effect of rehabilitation on speech outcomes," *Journal of Oral and Maxillofacial Surgery*, vol. 75, no. 7, pp. 1530–1541, 2017.

[44] I. Jacobi, M. A. van Rossum, L. van der Molen, F. J. Hilgers, and M. W. van den Brekel, "Acoustic analysis of changes in articulation proficiency in patients with advanced head and neck cancer treated with chemoradiotherapy," *Annals of Otology, Rhinology & Laryngology*, vol. 122, no. 12, pp. 754–762, 2013.

[45] Q. Wang, H. Muckenhirn, K. Wilson, P. Sridhar, Z. Wu, J. R. Hershey, R. A. Saurous, R. J. Weiss, Y. Jia, and I. L. Moreno, "VoiceFilter: Targeted Voice Separation by Speaker-Conditioned Spectrogram Masking," in *Proc. Interspeech 2019*, 2019, pp. 2728–2732.

[46] M. Illa, B. M. Halpern, R. van Son, L. Moro-Velazquez, and O. Scharenborg, "Pathological voice adaptation with autoencoder-based voice conversion," in *Proc. 11th ISCA Speech Synthesis Workshop (SSW 11)*, 2021, pp. 19–24.

4

AUTOMATIC EVALUATION OF SPONTANEOUS ORAL CANCER SPEECH USING RATINGS FROM NAIVE LISTENERS

In this paper, we build and compare multiple automatic speech systems for the automatic evaluation of the severity of a speech impairment due to oral cancer, based on spontaneous speech. To be able to build and evaluate such systems, we collected a new spontaneous oral cancer speech corpus from YouTube consisting of 124 utterances rated by 100 non-expert listeners and one trained speech language pathologist, which we made publicly available. We evaluated the speech severity on the level of single utterances, and on the level of full recordings. The results of extensive experiments showed that (1) the highest correlation with the human severity ratings were obtained with two automatic speech recognition (ASR) systems on the level of single utterances, and a modulation spectrum-based LASSO model on the level of recordings. (2) The use of binary labels led to lower correlations with the human ratings than using intelligibility scores. On the other hand, we found that naive listeners' ratings are highly similar to the speech pathologist's ratings for speech severity evaluation.

4.1. INTRODUCTION

Oral cancer is a type of cancer where a tumour is located inside the oral cavity, most typically the tongue or floor of the mouth. Approximately 530,000 people get diagnosed

This chapter has been submitted as: Halpern, B. M., Feng, S., van Son, R., van den Brekel, M., & Scharenborg, O. (2022). Automatic evaluation of spontaneous oral cancer speech using ratings from naive listeners. Speech Communication. The PhD candidate contributed to the data collection, implementation, experimental design, writing and evaluation of the experiments.

with this condition every year worldwide [1], including 53,000 in the USA [2]. To treat oral cancer, (part of) the tissues surrounding the tumour are removed during an operation, which subsequently affects the speech of the oral cancer patients. In certain cases, patients are able to learn articulatory compensation techniques to adjust for the lost tongue tissue [3]. Learning these compensation techniques as part of speech therapy can alleviate speech problems in oral cancer speakers.

To evaluate the success of such speech therapy, an automatic objective speech evaluation approach would be highly useful. Currently, speech-language pathologists (SLPs) are relying on standardised questionnaires such as the Grade-Roughness-Breathiness-Astenicity-Strain Scale (GRBAS) [4] and the Consensus Auditory Perceptual Evaluation of Voice (CAPE-V) [5] to evaluate the speech of their patients. However, these perceptual evaluation approaches are heavily criticised as they are subjective and sensitive to several confounds such as the type of stimuli (sentences, sustained vowels), environmental noise or the type of microphone/hardware used [6].

There are only a few existing methods that propose automatic evaluation for oral cancer [7, 8]. Moreover, these have only been validated with clean data and read speech. Training with clean data and read speech is not necessarily ecologically valid, i.e., using spontaneously elicited speech is certainly more representative of a patient's everyday speech, therefore, more indicative of the actual speech severity [9, 10]. Furthermore, an ideal objective evaluation method should be insensitive to channel noises and the type of recording devices.

Towards our aim to develop a more ecologically correct, robust objective evaluation method for oral cancer speech, we collected an oral cancer speech dataset from YouTube with a wide variety of realistic speech conditions, which is more representative of oral cancer speakers' everyday speech than a read speech corpora. To the best of our knowledge, this corpus, which is an extension of our previous oral cancer dataset [11], is the first publicly available oral cancer speech evaluation dataset. Two other datasets exist, a French and a German dataset; however, these are not publicly available [7, 8].

The speech severity evaluation task can be roughly described as a speech processing task, where either one or multiple speech signals are fed into a processing function to obtain a single scalar number ($\hat{x} \in \mathbb{R}$) which is the estimate of the speech severity. This estimate can be compared against a ground truth severity score (x), which is either obtained from a speech language pathologist or a naive human listener. The estimated speech severity is then correlated with the ground truth severity score, where a correlation of 1 indicates the perfect method for speech severity estimation.

The main aim of this work is to compare existing and new techniques for the automatic evaluation of the severity of the speech impairment due to oral cancer treatment (in short, oral cancer speech) to find the system that achieves the highest correlation with human ratings of the severity of the oral cancer speech. Therefore our main research question is the following: **RQ1: What automatic approach achieves the highest correlation with the ground truth severity scores for oral cancer speech severity evaluation?**

There are several paradigms for the objective evaluation of pathological speech. We divide these paradigms into two groups, i.e., reference-based and reference-free approaches. Reference-based methods use either a transcription of a speech signal (ASR-based meth-

ods) or a reference speech signal (comparison-based methods), while reference-free methods do not. In this work, we will compare both kinds of reference-based and several reference-free methods on the task of oral cancer speech severity evaluation.

ASR-based methods [12, 7] use the mistakes of speech recognisers to assess the speech quality of patients. In other words, it is assumed that an ASR makes similar errors as an expert. Some transcription error measure (e.g., phoneme error rate, word error rate, Levenshtein distance) is used as the severity estimate $\hat{x}$. ASR-based methods are often deemed as the most useful methods because practitioners can directly inspect what words or phonemes ASR systems did not recognise. Their main disadvantage, though, is that a ground truth transcription of the pathological speech is required, which is often difficult to obtain, especially when the speech is unintelligible. In this work, we are going to test several ASR-based techniques for oral cancer severity estimation from [13].

Comparison-based methods measure the distortion of a speech signal compared to a reference speech signal. These approaches originate from the speech enhancement (blind source separation) literature, where the distorted signal is a noised signal, which is compared to a clean signal [14]. Pathological speech, then, can be seen as a distortion of the healthy speech signal. An often used distortion measure in speech enhancement is the Short Time Objective Intelligibility method (STOI), and its variant ESTOI [15]. STOI is not directly applicable to pathological speech, as STOI assumes that the distorted (here: pathological) signal and the reference signal have equal duration, which is seldom the case. [16] proposed a modification of STOI and E-STOI, called P-STOI and P-ESTOI, which performs time alignment of the pathological and reference signals, and which can estimate severity with a high correlation to listener scores for two separate databases of dysarthric speech. Therefore, we include P-STOI and P-ESTOI in our comparison.

Recognising that advancements in speech enhancement evaluation can be applied to the evaluation of speech severity, we are interested if we can also apply techniques used in synthetic speech evaluation to oral cancer speech severity evaluation. Specifically, we investigate whether the most common objective approach used in synthetic speech evaluation, the Mel-cepstral distortion (MCD), can be used for the oral cancer speech severity estimation task [17].

Reference-free methods perform objective evaluation without the need for a transcription of the pathological speech signal or the need for a reference (healthy) speech signal. Instead they use a statistical model (e.g., a deep neural network or a LASSO model) and a feature representation to provide the severity estimate $\hat{x}$. We investigated the following possible features: (1) long-time average spectrum (LTAS), which has been used as a voice quality measurement in the detection of pathological speech [18, 19] and for the evaluation of the effect of speech therapy or surgery on voice quality [20]. Moreover, in our previous studies, LTAS was successfully used to differentiate between oral cancer speech [11] and healthy speech; and dysarthric speech and healthy speech [21]. (2) Speaker embeddings, which have attracted a lot of attention recently (i-vector [22, 23], x-vector [8], d-vector [24]), and seems to be useful for oral cancer speech intelligibility estimation [8]. (3) Moreover, we investigate how reference-free synthetic speech evaluation methods perform on the severity evaluation task, i.e., global variance (GV) [25] and modulation spectrum (MS) [26]. We will compare each feature using a LASSO-based statistical model. The LASSO model is used to predict the severity measure $\hat{x}$ from

the feature representation after training on the ground truth severity scores. We believe that using LASSO allows for a (1) fairer comparison of features than neural networks, where performance might be dependent on tuning, initialisation seeds, or the chosen network architecture, (2) and it is an explainable machine learning technique which is a common requirement in clinical practice.

Both the reference-free and the reference-based approaches need large amounts of training data. Along with the reference transcriptions mentioned before, ASRs require large amounts of speech data, which are not available for all languages. Comparison-based approaches require a reference healthy speech signal. On the other hand, reference-free approaches also require some form of human labelling, namely, the judgement of severity from the listeners. These resources are typically difficult to obtain. Therefore, it is important to consider whether we can reduce the cost of labelling. The secondary research question of the work is the following: **RQ2: Are other approaches available that require less labelled training data while giving similar performance on the speech evaluation task?**.

We investigate two possible approaches: (1) Instead of predicting the severity directly, we could predict the probability of absence/presence (classification/detector task) of oral cancer speech, after which this probability can be correlated with the severity scores. This classification/detector task only needs binary labels, which are substantially easier and cheaper to acquire as no expert annotators are needed. In other words, we are interested in whether detectors can achieve comparable performance to regressors. **(RQ2.1)**. (2) We propose to use the intelligibility ratings from naive listeners instead of expert listeners. To that end, we investigate in how far ratings from naive, non-expert listeners recruited through a crowdsourcing platform agree with those of expert listeners **(RQ2.2)**.

The paper is organised as follows. In Section 4.2, we explain how we gathered the oral cancer dataset used in this research, and we perform an initial exploratory analysis on the reliability of the collected ratings. The section ends with a comparison of naive and expert listeners where we answer **RQ2.2**. Section 4.3 explains the experimental design to answer the research questions and includes a methodological summary for each technique. Finally, Section 4.4 presents and discusses the results from the perspective of each research question. The dataset in this paper and the evaluation recipes are publicly available[1].

4.2. DATASET COLLECTION AND ANALYSIS OF THE RATING STUDY

The following sections present the oral cancer database, its collection and the oral cancer speech severity rating by naive listeners obtained through crowd-sourcing and by speech language pathologists (SLPs). This will be followed by an exploratory analysis of the collected ratings, which aims to investigate the reliability of the ratings. Moreover, we will answer research question (**RQ2.2**) whether the severity scores from naive listeners are comparable to those of speech language pathologists.

Table 4.1: Partitioning of the speakers into the training and evaluation set. RF stands for reference-free, RB stands for reference-based. The red colour indicates female speakers, while the blue colour indicates male speakers. The column "Phonetic cover(age)" indicates the percentage of the different phonemes in the lexicon (CMUDict) that is present in the utterance by that speaker. The column "VoxCeleb control" contains the id of the control speaker from the VoxCeleb dataset, which is used only during the detection task. In the case of the reference-free models, scores are extrapolated (see Section 4.3.2) and trained with all available audio, therefore the number of rated utterances (parentheses) differ from the number of utterances used for training.

Speaker	Training RF	Training RB	Evaluation RF	Evaluation RB	Utterances included	Phonetic cover	VoxCeleb control
id001			✓	✓	10	79.49%	
id002	✓				8	Unintelligible	id10571
id003	✓	✓			8	82.05%	id10078
id004	✓				8	Unintelligible	id10111
id005			✓	✓	10	94.87%	
id007	✓	✓			8	87.18%	id11250
id008		✓	✓		8	92.31%	
id010			✓	✓	3	74.36%	
id011	✓	✓			8	92.31%	id10242
id013			✓		10	Unintelligible	
id014			✓	✓	2	87.18%	
id015			✓	✓	3	74.36%	
id016			✓	✓	10	84.62%	
id017			✓	✓	10	92.31%	
id018			✓	✓	10	71.79%	
id019			✓	✓	8	84.62%	
Total speakers (16)	5	4	11	9	-	-	-
Total utterances (124)	1632 (40)	636 (32)	84	66	-	-	-
Total audio used	2h 16 min	1h 46 min	54 min	7 min	-	-	-

4.2.1. COLLECTION OF THE DATASET

We manually collected 3 hours of audio data containing English oral cancer speech from YouTube. The dataset includes 16 speakers. The presence of oral cancer speech was determined by the content of the video and the authors' (B.H., R.V.S., M.v.d.B.) clinical experience with such speakers. The audio was then manually cut to exclude music, healthy speakers, non-American English speakers. All utterances were downsampled to 16 kHz, loudness normalised to -0.1 dB, and finally mixed from stereo to mono using the `sox` tool. Transcriptions were created manually starting from baseline transcriptions generated by the Baseline ASR system explained in Section 4.3.3.

We distinguish the utterances based on whether the annotator (B.H.) was able to transcribe the utterance (intelligible) or not (unintelligible). The unintelligible utterances will only be used for the reference-free techniques. After preprocessing and splitting, the dataset contains a total of 840 transcribed 10-sec (140 min) long utterances, and an additional 936 5-sec long utterances (78 mins) of speech that is not transcribed. The dataset is partitioned into four different sets: a training and an evaluation set for both approaches (Reference-based and Reference-free). The reference-based evaluation set consists of the transcribed (intelligible) utterances. The reference-free approaches are also evaluated on the reference-based evaluation, to compare all approaches once using the same test set. The reference-free approaches are also evaluated on the reference-based evaluation set to investigate their effectiveness in a condition where a reference might not be available. Table 4.1 provides the details of the training and evaluation sets such as the amount of audio used and the number of utterances. The selection of speakers in the reference-free and reference-based approaches follow the setup used in our

[1] https://karkirowle.github.io/oral_cancer_corpus/

previous papers [13, 11].

4.2.2. SELECTION OF STIMULI FOR QUESTIONNAIRE

In order to determine which approach works best for the oral cancer speech severity evaluation task, we need ground truth ratings. Because it would be too costly to get ratings for all oral cancer speech utterances, we selected a subset of the oral cancer speech utterances for rating by the naive listeners and the expert listener.

The subset of utterances for rating was created by adhering to the following:

1. (whenever possible), of the speakers in the evaluation set, 10 utterances will be rated;
2. (whenever possible), of the speakers in both training sets, 8 utterances will be rated
3. sentences are selected such that they cover the highest number of different phonemes for each speaker (*phonetic coverage*);
4. if there are multiple recordings available for a given speaker, at least one utterance from each recording is used to maximise channel variability for the speaker (*recording diversity*). It is important that a recording is the whole stretch of speech recorded at once, it is not the same thing as an utterance, i.e., a recording can have multiple utterances.

Please note that the recordings that do not have transcriptions, cannot be optimised for phonetic content. These recordings are manually cut without taking phoneme coverage into account. On the other hand, the recordings with transcriptions are optimised for phonetic content. In order to do so, first, all the words in the utterances were mapped to ARPABET phonemes (stress markers were ignored) using the CMU Dictionary[2]. In order to select for each speaker the set of utterances that has the largest phonetic coverage, a greedy algorithm is used to obtain an approximate solution in each step. The greedy algorithm selects the utterance which maximises the loss function:

$$\mathscr{L}(A, B, \text{new}) = |A \setminus B| + \alpha \cdot \mathbb{I}_{\text{new}},$$

where A is the set of phonemes in an utterance and B is the set of already covered phonemes. In other words, the difference of the number of elements (cardinality) in each set is calculated at each step to obtain the new phonemes. The parameter $\alpha \in \mathbb{R}^+$ is a hyperparameter, which can be tuned for each speaker separately. $\mathbb{I}_{\text{new}}$ is an indicator function, which takes on the value 1 if the recording is new, otherwise it is 0. This parameter controls the importance of new recordings over the importance of new phonemes: an $0 < \alpha \leq 1$ means: given an equal number of new phonemes in two different candidate utterances, the utterance coming from a new recording is preferred. Extending this logic, we can see that for any arbitrary α where α is $k \leq \alpha \leq k+1$ ($k \in \mathbb{Z}^+$), the increasing of α allows for losing k additional phoneme(s), if the selected recording is new in the other candidate utterance. For most speakers, we could obtain a selection that has all

[2] http://www.speech.cs.cmu.edu/cgi-bin/cmudict

the recordings with $\alpha = 0.1$, in other words, there is no trade-off between the phonetic coverage and the recording diversity, with the exception of one speaker (id011), where we used $\alpha = 1.1$, this is equivalent of losing one phoneme.

The final selection for rating by the naive and expert listeners consisted of 98 intelligible utterances and 26 unintelligible utterances. These 124 utterances are henceforth referred to as "the stimuli".

4.2.3. DISTRIBUTION OF QUESTIONNAIRES

The rating study was administered via Qualtrics[3] and distributed through Prolific[4]. The 124 utterances were randomly assigned to one of five questionnaires (with one questionnaire containing only 24 utterances). Each of the 100 naive participants was then randomly assigned one of the questionnaires. Each participant could participate more than once, up to a maximum of five times. The stimuli in each questionnaire were randomised for each participant in order to average out possible learning effects by the participants. We found only one American English expert SLP to rate the audio samples, which further emphasises the need for automatic evaluations. The SLP rated all 124 utterances using the same method as the naive listeners.

The task for each participant (both the naive and expert listeners) was to rate the severity of each utterance on a 5-point Likert scale. This 5-point rating scheme is quite different from standard SLP evaluations mentioned in the introduction, however, we wanted to design a task that was not too difficult for naive listeners to carry out. Therefore, we refrained from evaluating qualities such as breathiness, hoarseness and using a visual analogue scale, which is often part of standardised evaluations done by SLPs. Each questionnaire started with an example of a completely healthy utterance taken from the CMU Arctic corpus [27], and an example of a pathological utterance taken from the TORGO [28] corpus.

4.2.4. RESULTS OF THE NAIVE LISTENER RATING STUDY

First we assessed the consistency of the ratings for each speaker and the global tendency of the rating, i.e., an often-heard criticism of uneven Likert scales is that the "neutral" (3 in our case) score is used as a fall-back option and hence has the highest frequency.

To assess if there is any global tendency within the results, we looked at the mean scores for each recording in Table 4.2 and the histogram of the dataset, see Figure 4.1. The lowest mean score was for recording 8 (1.05), while the highest was for recording 18 (4.90), which indicates that participants used the full extent of the rating scale. Furthermore, the histogram on Figure 4.1 shows that a rating of 5 (healthy speech) was most commonly used. It is true that the obtained ratings do not seem to exhibit a completely uniform distribution, but this is more likely due to the fact that the severity of the utterances were not controlled when selecting the utterances.

We then carried out an analysis of the range of the means of the ratings of the recordings per speaker. The upper part of Table 4.2 lists all the mean scores for each recording, grouped by speakers. There are 5 speakers (id001, id002, id004, id008, id011) who have

[3] https://www.qualtrics.com/
[4] https://www.prolific.co/

Table 4.2: Mean ($\bar{x}$) and standard deviation (s) of naive listener scores (top) and speech language pathologist scores (bottom) obtained for each recording (multiple utterances are rated for each recording) and speaker rated in the rating study. Spk stands for the speaker id, Rec stands for the recording id.

Naive listener scores

Spk	id001		id002						id003	id004				id005	id007	id008
Rec	1	3	8	12	14	16	19	25	10	11	15	27	29	18	21	23
$\bar{x}$	3.03	4.28	1.05	1.13	1.38	1.16	1.12	1.79	1.98	1.18	1.06	1.11	1.08	4.90	4.26	2.39
s	0.78	0.75	0.22	0.34	0.60	0.56	0.47	0.59	0.66	0.38	0.24	0.42	0.27	0.33	0.74	0.80
Spk	id008	id010	id011							id013	id014	id015	id016	id017	id018	id019
Rec	24	31	4	5	6	7	13	22	28	17	30	32	33	34	35	36
$\bar{x}$	2.60	4.21	4.06	4.08	4.66	4.21	4.31	3.73	3.59	1.29	4.57	3.28	3.84	4.75	2.44	3.90
s	0.75	0.73	0.76	0.72	0.53	0.78	0.74	0.95	0.83	0.48	0.57	1.09	0.86	0.67	0.71	0.75

Speech language pathologist (SLP) scores

Spk	id001		id002						id003	id004				id005	id007	id008
Rec	1	3	8	12	14	16	19	25	10	11	15	27	29	18	21	23
$\bar{x}$	3.5	4.5	1.0	1.0	1.5	1.0	1.0	1.0	2.24	1.0	1.0	1.0	1.0	5.0	3.91	2.33
s	0.5	0.5	0.0	0.0	0.5	0.0	0.0	0.0	0.65	0.0	0.0	0.0	0.0	0.0	0.68	0.47
Spk	id008	id010	id011							id013	id014	id015	id016	id017	id018	id019
Rec	24	31	4	5	6	7	13	22	28	17	30	32	33	34	35	36
$\bar{x}$	2.6	4.67	5.0	5.0	5.0	5.0	5.0	4.0	5.0	1.2	4.5	4.0	4.40	4.9	2.8	4.0
s	0.8	0.47	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.4	0.5	0.0	0.49	0.3	0.6	0.0

multiple recordings. Three of them have a score range of more than 0.5: id001 (range = 1.25), id011 (range = 1.07), id002 (range = 0.74). In the case of id002, there are two recordings (Recording 14 and 25) that seem to receive noticeably higher scores (1.79 and 1.38) than the other recordings of the speaker. Moreover, a Wilcoxon signed-rank hypothesis test (see Table 4.6 for the p-values) showed a significant difference between the distribution of scores for the recordings of id008. Finally, in the case of id004, recording 11 was significantly different compared to recording 15, but otherwise, the ratings were consistent.

These differences in ratings for recordings by the same speaker can likely be explained by differences in the time when the recordings were created rather than inconsistencies in the ratings by the naive listeners. For example, speakers id011 and id001 self-report that their videos were recorded at different moments in time, where they have a different speech severity, which might explain the rather large range for these two speakers. In the case of id002, informal listening by author B.H. confirmed that recordings 14 and 25 indeed were a lot more intelligible than the other recordings from speaker id002. We hypothesise that this is because the recordings were done at a different time, however, this is not obvious from the content nor the corresponding metadata of the recordings. Furthermore, the scores seem to be well aligned with the scores of the expert listener, see Section 4.2.5.

4.2.5. RQ2.2: COMPARISON OF NAIVE AND EXPERT LISTENERS

In order to compare the naive and expert listener scores, we used a Pearson's correlation between the mean of the scores from the naive listeners and the mean of the scores from the SLP. The strength of the correlation was $r = 0.92$ ($p < 0.001$), which is very high. This strongly indicates that we can use ratings from naive listeners obtained through crowd-

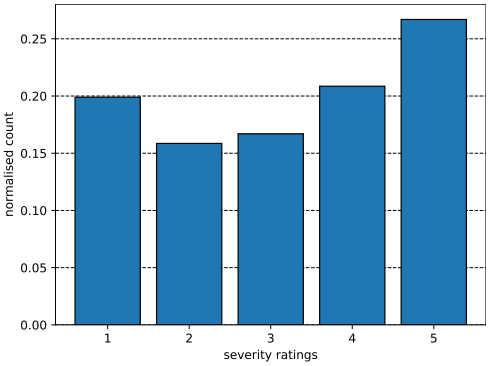


Figure 4.1: Histogram of all the ratings in our dataset. The x-axis shows the ratings given, 1 being the most severe, 5 being the least severe or healthy, the y-axis shows the normalised counts.

sourcing to rate the severity of the speech reliably on a 5-point scale. For the rest of this paper, we will use the naive listener scores as ground truth severity scores to validate our different methods, as these scores are based on more raters and are thus more granular than that of a single SLP.

Please note, in general, we expect that differences between the ratings of the naive listener and SLP would be much more apparent in evaluation questionnaires that ask for explicit speech qualities such as breathiness, (i.e., as in GRBAS [6]), as naive listeners have little understanding about the acoustic cues corresponding to these terminologies.

4.3. METHODS

4.3.1. EXPERIMENTAL DESIGN

Table 4.3 lists all models that were compared in order to find the best technique for oral cancer speech severity evaluation. For each model, we indicate whether it uses a reference (see column "Reference"), and if so, which type of reference (column "Reference type"). For the ASR reference-based experiments (Baseline, Baseline+OC, DNN for AM Retraining, fMLLR), there are additional variants that we have not listed in the table for the sake of clarity, please see Section 4.3.3 for more details.

All models will be compared on the reference-based evaluation set, while the reference-free models will also be compared on the reference-free evaluation set. In order to find the best technique for oral cancer speech severity evaluation, the intelligibility estimate $\hat{x}$ obtained for each model is correlated with the average severity rating obtained from the naive listeners.

4.3.2. REFERENCE-FREE APPROACHES

To evaluate the reference-free approaches in a consistent way, we will use a LASSO-based detection and regression model (Section 4.3.2). The LASSO model will be tested with the d-vector (dvec), x-vector (xvec)(Section 4.3.2), LTAS (Section 4.3.2), and the global variance and the modulation spectrum (Section 4.3.2) features.

Table 4.3: Overview of all systems evaluated in this paper. ✗means reference-free, ✓means reference-based. * means that the synthetic reference is only used in the SynthNorm variants. Reference-free models are trained with the reference-free dataset, and reference-based models are trained with the reference-based dataset.

Model	**Reference**	**Reference type**
GV-detector	✗	No
GV-regressor	✗	No
MS-detector	✗	No
MS-regressor	✗	No
LTAS-detector	✗	No
LTAS-regressor	✗	No
xvec-detector	✗	No
xvec-regressor	✗	No
dvec-detector	✗	No
dvec-regressor	✗	No
Baseline	✓	Transcription, synthetic*
Baseline+OC	✓	Transcription, synthetic*
DNN for AM Retraining	✓	Transcription, synthetic*
fMLLR	✓	Transcription, synthetic*
MCD	✓	Synthetic
P-STOI	✓	Synthetic
P-ESTOI	✓	Synthetic

REFERENCE-FREE EXPERIMENTAL SETUP

LASSO is a variant of linear regression [29], which performs feature selection and regression simultaneously. Potentially, for a given linear regression task, some features do not contain any relevant information to make predictions or contain information that is collinear with the other features, causing overfitting. In LASSO, coefficients of regression are encouraged to be close to zero if they do not provide useful information. Zeroing (pruning) some features means that the model requires only a subset of all predictors, making the statistical model parsimonious and easier to interpret.

There are two variants of LASSO that we will use, one for regression and one for detection. For regression, we will use the vanilla LASSO. At inference time, vanilla LASSO's computation is identical to linear regression (Equation 4.1), however, at training time the coefficients ($\mathbf{w} \in \mathbb{R}^m$ where m is the dimensionality of the feature) are obtained in a slightly different way by adding the sparsity penalty to the ordinary least squares loss function (see Equation 4.2):

$$\hat{x}_i = \mathbf{w}^T \mathbf{h_p}(i), \tag{4.1}$$

$$\hat{\mathbf{w}} = \text{argmin}_{\mathbf{w}} ||\mathbf{w}^T \mathbf{h_p}(i) - x||_2^2 + \lambda ||\mathbf{w}||_1. \tag{4.2}$$

Pruning of the features is facilitated by setting the parameter $\lambda = 0.1$: the larger this parameter is, the closer the coefficients are to zero.

For detection, we will use a logistic LASSO, which is similar to LASSO with two key differences: (1) the addition of the sigmoid function to obtain the detection probability; (2) instead of x, binary labels are used, which we denote with $x_b \in \{0, 1\}$.

$$\hat{x}_i = \sigma(\mathbf{w}^T \mathbf{h_p}(i)) \quad \sigma(x) = \frac{1}{1 + \exp(-x)}, \tag{4.3}$$

$$\hat{\mathbf{w}} = \text{argmin}_{\mathbf{w}} ||\sigma(\mathbf{w}^T \mathbf{h_p}(i)) - x_b||_2^2 + \lambda ||\mathbf{w}||_1. \tag{4.4}$$

Note that this model is effectively a perceptron with L_1 regularisation. The addition of the sigmoid function does not cause any problems with the optimisation, as the sigmoid function is differentiable with respect to $\mathbf{w}$.

In order to compute the different reference-free features, we first chunk the utterances into 5 seconds segments ($\mathbf{y}_p(i) \in \mathbb{R}^T$, where i is the chunk index, and T is the total number of chunks) – note that the last chunk's duration can be shorter than 5 seconds. Subsequently, different features are extracted (see the later sections in Section 4.3.2) for each of these 5-sec chunks, where we obtain $\mathbf{h}_p(i) \in \mathbb{R}^d$, where d depends on the kind of feature we are using, and i denotes the chunk index. Therefore a training pair consists of the chunk of the recording $\mathbf{h}_p(i)$, and the corresponding severity score x.

The final prediction scores are obtained slightly differently for the reference-based and for the reference-free evaluation set. For the reference-based evaluation, we use information exclusively from the rated utterances. In this case, the final severity estimate is simply the mean of all chunk estimates in a single utterance,

$$\hat{x} = \frac{1}{T} \sum_{i=1}^{T} \hat{x}_i \tag{4.5}$$

For the reference-free evaluation, we cannot setup the experiment so it is comparable with the reference-based evaluation. Therefore, we decided to use a different, recording-level evaluation setup for the reference-free evaluation. There are two advantages of this setup over the utterance-level setup. First, the recording-level evaluation setup is a more sound setup from a clinical linguistic perspective, as this approach takes into account the fact that the impact of oral cancer surgery on pronunciation is different for different sounds. Therefore, the perceived severity should be also different for the different parts of the recordings as they contain different sounds. Second, the recording-level setup is able to use more data for the severity estimation than the utterance-level setup.

Recording-level evaluation creates a recording-level score $\hat{x}$ from all the available utterance chunks within a recording. In other words, the main difference is that all the chunks are considered in the recording, not only the chunks of a single utterance. In the recording-level evaluation, the final score is obtained as the weighted average of the scores for each utterance, i.e., for a recording that has n utterances with a number of chunks T_1, T_2 and T_n:

$$\hat{x} = \frac{1}{T_1 + T_2 + ... + T_n} \left(\sum_{i=1}^{T_1} \hat{x}_{1,i} + \sum_{j=1}^{T_2} \hat{x}_{2,i} ... + \sum_{k=1}^{T_n} \hat{x}_{n,k} \right). \tag{4.6}$$

Each reference-free LASSO model was trained with the Reference-free training set, which includes both the intelligible and the unintelligible utterances. As only a selection of the utterances in the corpus, and thus of the reference-free training set, was rated by human listeners, we extrapolated these ratings for those utterances without ratings in order to increase our training set size. All utterances without a rating received the average rating calculated over all rated utterances of the same recording of that speaker. The extrapolated ratings were also used as ground truth ratings, and are referred to as x.

SPEAKER EMBEDDINGS

The two speaker embedding features tested in this work are the angular x-vector (xvec) (which is an improved version of the x-vector) and the d-vector (dvec) [30, 31]. To extract a speaker embedding, the $\mathbf{y}_p(i)$ was fed through a deep neural network (DNN). Instead of the class labels, the activations of one of the intermediate layers were extracted and used as the speaker embeddings feature $\mathbf{h}_p(i)$ in our LASSO model.

Angular x-vectors differ from the conventional x-vector model [32] by using an angular softmax function instead of the normal softmax function, and using SincNet features instead of MFCCs [33]. The d-vector uses the generalised end-to-end loss (GE2E) as its loss function, while having 40-dimensional Mel filterbank features. For both of these models, we used (previously) publicly available implementations[5].

LTAS

In order to obtain the LTAS features, we extracted a (so-called) Kaldi spectrogram from the audio chunk $\mathbf{y_p}(i)$ with a 25 ms length Povey window, 10 ms frame shift and 256 frequency bins using the PyTorch torchaudio library. The obtained spectrogram is denoted

[5] https://huggingface.co/hbredin/SpeakerEmbedding-XVectorMFCC-VoxCeleb.
https://github.com/resemble-ai/Resemblyzer

by $\mathbf{S}_p \in \mathbb{R}^{256\times L}$, where 256 is the number of frequency bins and L is the number of analysis frames in the spectrogram, which is dependent on the duration of the individual chunks. We obtain the LTAS vector by stacking the mean and standard deviation for all 256 frequency bins which results in a $\mathbf{h}_p \in \mathbb{R}^{512}$ LTAS vector:

$$\mathbf{h}_p = \begin{bmatrix} \frac{1}{L}\sum_{j=0}^{L}\mathbf{S}_p(0,j) \\ \vdots \\ \frac{1}{L}\sum_{j=0}^{L}\mathbf{S}_p(255,j) \\ \sqrt{\frac{\sum_{j=0}^{L}\mathbf{S}_p(0,j)-\mathbf{h}_p(0)}{L-1}} \\ \vdots \\ \sqrt{\frac{\sum_{j=0}^{L}\mathbf{S}_p(255,j)-\mathbf{h}_p(255)}{L-1}} \end{bmatrix} \tag{4.7}$$

GLOBAL VARIANCE AND MODULATION SPECTRUM

Both the global variance (GV) and modulation spectrum (MS) are commonly used to evaluate synthetic speech objectively. For the GV calculation, we first calculated 20-dimensional librosa MFCC trajectories ($\mathbf{c}_p(i) \in \mathbb{R}^{20\times M}$) from the audio chunks $\mathbf{y}_p(i)$. From each MFCC trajectory, we calculated a time-axis variance estimate, which resulted in the 20-dimensional GV features, using:

$$\mathbf{h}_p(i) = \frac{1}{M}\sum_{j=1}^{M}\mathbf{c}_p(i)(j) - \bar{\mathbf{c}_p}(i) \quad \bar{\mathbf{c}_p}(i) = \sum_{j=1}^{M}\mathbf{c}_p(i)(j). \tag{4.8}$$

For the MS, we used the implementation from nnmnkwii [34]. First, we extracted 60-dimensional Mel-generalised cepstrum coefficients (MGC), and ignored the 0th order MGC. Subsequently, we took the power of the discrete Fourier transform of the MGC parameter trajectory across the time-axis. To obtain a duration-independent feature, we took the time-axis average, which resulted in the final 59-dimensional MS feature, which was computed using:

$$\mathbf{h}_p(i) = \frac{1}{M}\sum_{i=1}^{M}(\mathscr{F}\{\mathbf{c}_p\})^2(i). \tag{4.9}$$

4.3.3. REFERENCE-BASED

WORD-LEVEL ASR SYSTEMS

We used four different ASR systems from our previous work [13] to generate word-level transcriptions for oral cancer speech recordings: (1) Baseline (see Section 4.3.3), (2) Baseline + oral cancer (Baseline + OC; see Section 4.3.3), (3) DNN for AM retraining (see Section 4.3.3), and (4) feature-space maximum likelihood linear regression (fMLLR) based system (see Section 4.3.3). All four ASR systems were trained by leveraging data from Wall Street Journal (WSJ) speech corpus (healthy speech) and oral cancer speech except the baseline system, which was trained only on WSJ speech. For each system, we created a variant with a tri-gram language model and an RNN (LM). Furthermore, for

each language model-system pair, we ran a variant that uses synthetic speech references (SynthNorm) motivated by the work of [35] (see SynthNorm in Section 4.3.3).

The Levenshtein distance has previously been found to perform well for speech severity evaluation using ASR systems [12]. Therefore, here we used this same measure. The Levenshtein distance was calculated between the ground truth transcription (with the exception of SynthNorm, see Section 4.3.3) and decoded transcription of each utterance, and subsequently correlated with the average rating from the naive listeners.

Baseline: The baseline system is a standard hybrid DNN-HMM ASR system which is trained exclusively on healthy speech using the `si284` set of the WSJ corpus [36]. The acoustic model (AM) of the baseline system consisted of 5 feed-forward hidden layers of dimension 1,500 and a softmax output layer of 3,431 (equal to the number of HMM states). The input features to the DNN AM were 23 dimensional filterbank plus 3 dimensional pitch features (FB+P). We followed the Kaldi recipe[6] in training the baseline DNN AM.

Baseline + OC: The system baseline + OC followed a similar training pipeline as the baseline system, with the exception of using both the WSJ `si284` data and the oral cancer training data to train the DNN AM.

DNN for AM retraining: The DNN for AM retraining system was based on the baseline system (in Section 4.3.3) and sequently retrained. Specifically, the baseline DNN-HMM AM was used to generate forced-alignments for the oral cancer training data using its reference transcriptions. Next, the oral cancer training speech and its corresponding alignments were taken as training data and labels to re-train the DNN-HMM AM.

fMLLR: The fMLLR system aimed at leveraging the success of the fMLLR algorithm in speaker adapted feature (named fMLLR feature) generation [37]. In the context of oral cancer speech recognition, the use of fMLLR features could suppress pathological speech sound characteristics in oral cancer speech, encouraging oral cancer speech representations to be more similar to those of normal speech, hence improving the recognition performance [13]. Similar to the baseline + OC system, the fMLLR system was trained using both the WSJ and oral cancer speech data, with the only difference being of applying fMLLR features (40 dimension) instead of FB+P features during DNN AM training. The fMLLR features were estimated during GMM-HMM training, also using the merged WSJ and oral cancer speech data.

SynthNorm normalisation:

Inspired by [35], who found that comparing dysarthric speech to text-to-speech output can result in robust intelligibility estimation, we are going to experiment with an additional normalisation step in calculating our Levenshtein distance which we are going to call SynthNorm. Instead of using the ground truth transcription directly, we will generate a reference speech sample using a text-to-speech synthesis (TTS) system, and then use an ASR to recognise the synthesised speech, which should again result in the ground truth transcription if all went well. The expectation is that this normalisation step will remove errors in the severity estimation which are consistently made on both pathological and healthy speech. The distinction between any errors in the estimation due to healthy speech and errors due to the pathology are important for us, as we are only interested in errors due to the pathological quality of the utterance. The synthesised speech is

[6] `kaldi/egs/wsj/s5`

generated using a highly natural Tacotron-2 text-to-speech synthesis (TTS) system[7] [38]. Next, all the ASR systems introduced in Section 4.3.3 are used to decode both the original pathological speech utterance (predicted transcription) and the synthesised version of the same utterance (reference transcription). The Levenshtein distance between the reference and predicted transcription is then calculated in the standard way.

COMPARISON-BASED APPROACHES

The comparison-based approaches require a reference, healthy speech signal ($\mathbf{y}_r \in \mathbb{R}^{d_r}$, where d is the duration of the reference signal) along with the pathological signal ($\mathbf{y}_p \in \mathbb{R}^{d_p}$ where d_p is the duration of the pathological signal). Because there are no healthy references available, we will use the synthetic speech references already described in Section 4.3.3.

P-STOI and P-ESTOI: Both P-STOI and P-ESTOI are modifications of the STOI technique, commonly used in the speech enhancement field. STOI does not account for the different tempi of healthy and pathological speech and assumes time-aligned speech signals. To account for the time-alignment issue, P-STOI and P-ESTOI extend the STOI technique with dynamic time warping (DTW). The calculation of the P-STOI/P-ESTOI scores is as follows. First, we extract the 1/3 octave band time-frequency (TF) representation $\mathbf{H}_p$ and $\mathbf{H}_r$ from $\mathbf{y}_r$ and $\mathbf{y}_d$, where we align $\mathbf{H}_r$ and $\mathbf{H}_p$ using DTW. We estimate the cross-correlation between the aligned representations. As these representations are two-dimensional, the cross-correlation can be done along either the temporal or the spectral axis. The temporal estimate is called the P-STOI score, while the spectral estimate is called the P-ESTOI score [16]. The estimated scores are used as our severity measure $\hat{x}$.

MCD: The Mel-cepstral distortion (MCD) metric is usually used to measure the difference between a synthetic and a natural speech signal in order to objectively evaluate synthesis quality in TTS development. Here, the MCD metric is used to measure the difference between the pathological speech signal $\mathbf{y_p}$ and the reference speech signal $\mathbf{y_r}$ in order to predict the severity score $\hat{x}$. To calculate the MCD, we first extracted 20-dimensional Mel frequency cepstral cepstral coefficients (MFCCs) from $\mathbf{x_p}$ and $\mathbf{x_r}$ using the librosa Python library [39]. We denote the obtained representations with $\mathbf{H_p} \in \mathbb{R}^{20\times M}$ and $\mathbf{H_r} \in \mathbb{R}^{20\times L}$ where L and M represent the number of analysis frames in the MFCC. The reference and pathological MFCCss have to be aligned if they have different length, otherwise calculation of the MCD is impossible. Therefore dynamic time warping (DTW) is performed to align the MFCCs. The aligned reference MFCC is denoted as $\mathbf{H_{rp}} \in \mathbb{R}^{20\times M}$. Following standard procedure, the α scaling coefficient was used [40]. Note that the zeroth-order MFCC is ignored following standard practice because it is dependent on the gain of the speech, which can be sensitive to noise.

$$\hat{x} = \text{MCD}(\mathbf{H_p}, \mathbf{H_{rp}}) = \frac{\alpha}{M}\sum_{i=1}^{M}\sqrt{\sum_{j=1}^{19}(\mathbf{H_p}(i,j) - \mathbf{H_{rp}}(i,j))^2} \quad \alpha = \frac{10\sqrt{2}}{\ln 2} \tag{4.10}$$

Table 4.4: Pearson's correlation of all the approaches evaluated on the reference-based evaluation set, rounded to two decimals. A cyan background colour marks the ASR acoustic models which use oral cancer data during training. "TTS reference" indicates whether a synthetic speech ground truth is used. *** indicate $p < 10^{-3}$, otherwise p-value is provided. The best performing model is emphasised with a **bold** typeface.

Reference-free approaches				
Model	**Pearson's r**	**p**	**Language model**	**TTS reference**
LTAS-detector	0.29	0.02	N/A	N/A
LTAS-regressor	0.39	0.001	N/A	N/A
dvec-detector	0.46	***	N/A	N/A
dvec-regressor	0.28	0.02	N/A	N/A
xvec-detector	0.32	0.007	N/A	N/A
xvec-regressor	0.34	0.005	N/A	N/A
GV-detector	0.32	0.009	N/A	N/A
GV-regressor	0.34	0.004	N/A	N/A
MS-detector	0.21	0.10	N/A	N/A
MS-regressor	0.45	***	N/A	N/A
Reference-based approaches (ASR-based)				
Baseline	0.56	***	n-gram	∅
Baseline + OC	**0.58**	***	n-gram	∅
DNN for AM Retraining	0.55	***	n-gram	∅
fMLLR	0.47	***	n-gram	∅
Baseline	0.56	***	RNN	∅
Baseline + OC	0.49	***	RNN	∅
DNN for AM retraining	0.57	***	RNN	∅
fMLLR	0.43	***	RNN	∅
Baseline	**0.58**	***	n-gram	Yes
Baseline + OC	0.57	***	n-gram	Yes
DNN for AM retraining	0.53	***	n-gram	Yes
fMLLR	0.45	***	n-gram	Yes
Baseline	0.55	0.001	RNN	Yes
Baseline + OC	0.50	***	RNN	Yes
DNN for AM retraining	0.57	***	RNN	Yes
fMLLR	0.43	0.002	RNN	Yes
Reference-based approaches (comparison-based)				
MCD	0.12	0.32	N/A	Yes
P-STOI	-0.02	0.85	N/A	Yes
P-ESTOI	0.14	0.27	N/A	Yes

Table 4.5: Pearson's correlation of the reference-free approaches on both the RB and RF evaluation sets. Of each detector/regressor pair, red background indicates a worse correlation while green indicates a better correlation than the other member of the pair. *** indicates p-values < 10^{-3}, otherwise p-value is written. The best performing model is emphasised with a **bold** typeface for each evaluation type (reference-free and reference-based). Note that the data in the right column of the table is identical to the top part of Table 4.4, we present the data twice for ease of understanding.

Reference-free approaches				
	RF evaluation (recording-level)		RB evaluation (utterance-level)	
Model	**Pearson's r**	**p**	**Pearson's r**	**p**
GV-detector	0.64	***	0.32	0.009
GV-regressor	0.72	***	0.34	0.004
MS-detector	0.68	***	0.21	***
MS-regressor	**0.76**	***	0.45	0.04
LTAS-detector	0.27	***	0.29	0.02
LTAS-regressor	0.66	***	0.39	0.0012
dvec-detector	-0.46	***	**0.46**	***
dvec-regressor	0.69	***	0.28	0.02
xvec-detector	0.55	***	0.32	0.008
xvec-regressor	0.53	***	0.34	0.005

4.4. RESULTS

4.4.1. RQ1: COMPARISON OF ALL APPROACHES ON THE SPEECH SEVERITY EVALUATION TASK

Table 4.4 lists the Pearson's correlations of the estimated severity score of all approaches with the average human rating of the naive listeners. All results are obtained on the reference-based evaluation set. The table is divided into three blocks. The upper part of the table shows the reference-free, the lower part of the table shows the reference-based approaches in two blocks: one block is for the ASR models, the other block includes the comparison-based approaches. When a model has a higher Pearson's correlation than another model, we will say that it outperforms the other model.

Comparing all approaches, we see that the Baseline+OC+ngram and the Baseline+ngram+TTS models performed the best on the reference-based evaluation set. This means that reference-based approaches seem to outperform reference-free approaches in determining oral cancer speech severity when a reference is available for evaluating the speech severity. We will further discuss the possible reasons between the performance differences of the ASR models in Section 4.4.3 (data differences), 4.4.4 (language model differences) and 4.4.5 (normalisation differences).

The reference-free approaches achieved moderate correlations with the average listener scores on the reference-based (utterance-level) evaluation set: the best approach was the dvec-detector, followed by the MS-regressor and the LTAS-regressor. We did not observe any obvious patterns in these results, so these will not be further discussed. Finally, most comparison-based approaches performed quite poorly on the reference-

[7] https://github.com/NVIDIA/tacotron2

based evaluation. We will discuss these results in Section 4.4.6.

Table 4.5 shows the results for the reference-free detector and regressor approaches on the reference-free evaluation set. The left column shows that the best approach on the reference-free evaluation set was the MS-regressor, followed by the GV-regressor and the dvec-regressor. (**RQ1**). For both the regression and the detection task, the best features are those that are used in the evaluation of synthetic speech. We will further discuss the general implications of this in Section 4.5. The speaker embeddings/features sometimes perform surprisingly poorly (dvec-detector r=-0.46), and sometimes perform comparable to the best methods (i.e., in the case of the dvec-regressor). It is as yet unclear why.

4.4.2. RQ2.1: CAN DETECTORS ACHIEVE COMPARABLE PERFORMANCE TO REGRESSORS ON THE SPEECH SEVERITY EVALUATION TASK?

Comparing the correlations of the regressors with the detectors on the reference-based evaluation set (right columns of Table 4.5) and reference-free evaluation set (left columns of Table 4.5), we observe that the regressors consistently achieved higher correlations than the detectors, with only two exceptions: the xvec on the reference-free recording level evaluation, and d-vec on the reference-based utterance-level evaluation. However, it is interesting to note that for the utterance-level evaluation, the best correlation with the human ratings was obtained for the dvec-detector . Overall (combining the recording-level and utterance-level evaluations) the regression experiment was better in 80% of the cases.

These results show that the regressor models which were trained on the intelligibility scores rather than the binary scores as the detectors were, are more informative for and better at the oral cancer severity evaluation task. Therefore, using binary class labels instead of intelligibility scores is not a good solution when one wants to build automatic methods to evaluate the severity of oral cancer speech that have a good correlation with human ratings of the severity of the oral cancer speech.

4.4.3. ORAL CANCER DATA SEEMS TO HELP IN ASR-BASED ORAL CANCER SEVERITY EVALUATION

From Table 4.4 we can see that the model that has the highest correlation with the human ratings is always a model that uses oral cancer data during training of the acoustic models (Baseline+OC, DNN AM Retraining) except in the case of the SynthNorm models using an RNN language model, where the Baseline is the best. We expect that adding some oral cancer data to the training material is beneficial to the ASR models because the acoustic models then capture some of the mild disfluencies due to oral cancer speech in a vein that is similar to how human listeners quickly adapt to mild disfluencies in healthy speech [41, 42]. It is interesting to note that even though the fMLLR uses oral cancer speech, it always achieves worse performance than the Baseline. We hypothesise that fMLLR adapts to the severity of the speaker and as such is able to learn the deviant pronunciations of an oral cancer speaker. Since fMLLR takes into account the entire recording of the speaker, this may result in an "overadaptation" to the oral cancer speaker. On the other hand, human listeners only hear a single utterance of a speaker at any given time, which is not enough to adapt to the deviant pronunciations of the oral

cancer speaker. Consequently, the scores provided by the fMLLR models do not correlate that well with the human ratings compared to the models without fMLLR.

4.4.4. WEAKER LANGUAGE MODELS SEEM TO LEAD TO IMPROVED CORRELATIONS

We can see that n-gram based language models outperformed the otherwise identical RNN-based models in nearly all cases except for the DNN for AM retraining models. A more complex language model thus does not generally improve the correlation with listener scores. This makes sense: a stronger language model (here the RNN) will help the ASR to decode the acoustic signal using stronger lexical and semantic information than a weaker LM. This means that a model that uses a stronger LM will rely less on acoustic cues, while these acoustic cues are more helpful for severity evaluation.

4.4.5. EFFECT OF SYNTHNORM ON THE RESULTS

When comparing the ASR models which use TTS and which do not use TTS reference, the SynthNorm models performed approximately on par. Two out of eight times the SynthNorm models performed better (Baseline+ngram, Baseline+OC+RNN), and two out of eight times the models performed equally well (DNN for AM retraining+RNN, fMLLR+RNN). Furthermore, a SynthNorm-based model (Baseline+ngram) was one of the best performing models. Therefore, we think that the presented SynthNorm approach is worth considering when doing severity evaluations, as it can improve the results in certain cases.

4.4.6. COMPARISON-BASED METHODS SEEM TO BE LACKING IN PERFORMANCE

In general, we can see that the comparison-based approaches, i.e., MCD, P-STOI, and P-ESTOI, performed poorly compared to the other approaches. We hypothesise that this might be due to the DTW, which is used in all of the comparison-based techniques. We think that the DTW might not be a robust aligner in the noisy conditions that are sometimes present in the dataset. Therefore, future work should look at other alignment methods (such as attention), and use multiple references to test their robustness under noisy conditions. It is likely that at least the P-STOI and P-ESTOI would improve when using multiple references as these methods are normally used with more references, however, that would have been an unfair comparison in the current study.

4.5. DISCUSSION

In this paper, we built and compared multiple automatic speech evaluation systems for the evaluation of the severity of a speech impairment due to oral cancer, based on spontaneous speech.

Our main research question concerned finding the best method for the automatic evaluation of oral cancer speech. The best method for the automatic evaluation of oral cancer speech was the modulation spectrum regressor, for which no reference transcription is needed. If reference transcriptions are available, then automatic speech recognisers can be used, which showed the highest correlation with the naive listener ratings on

the reference-based evaluation.

The majority of the methods showed a high to moderate correlation with the naive listener ratings, depending on the type of evaluation (utterance-level, recording-level) used. These scores are, however, considerably lower than one would normally find for clean, read pathological speech [16, 7, 8]. Our lower results are due to the spontaneous nature of the speech and the fact that the recordings were obtained from YouTube and contained substantial background noise. Therefore, for a clinical use case, we would still recommend using models based on read speech recorded with high quality microphones, such as the [8] uses.

In this paper, we also tested three methods that are traditionally used for synthetic speech evaluation[8]: the GV detector/regressor, the MS detector/regressor (tested both on recording-level and utterance-level) and the MCD (tested only on utterance-level). This is a very interesting repurposing of these methods as there are many commonalities between speech severity and speech naturalness, which has been studied only recently, as a clear distinction would help in evaluating synthetic pathological voices [21]. In our results, we found that the GV and the MS feature based models both performed very well in comparison to the other features that we have tested: On the recording-level, the MS-regressor had the highest correlation with the human ratings, and the GV-regressor had the second highest correlation. On the utterance-level task, however, they were often outperformed by the other tested methods. These results indicate that speech synthesis evaluation approaches are working well on a recording-level, but not on the utterance-level. We hypothesise that the quality of the speech ("naturalness") is an important aspect of the severity when evaluating on the recording-level, while on the utterance-level the intelligibility is much more important - this latter observation is further corroborated by the ASR-based models being the best on the utterance-level, which primarily capture the intelligibility information. Therefore, we think that future research efforts should be directed towards a deeper understanding of the boundary between naturalness and severity. A possible avenue would be to present stimuli of various speech severities to naive listeners, similar to our current study, however, with different levels of noise mixed into the stimuli, asking for not only the speech severity but also the naturalness of the utterances. Such a study would allow a detailed investigation of speech severity and naturalness simultaneously on the subjective level, which would enable further development of objective methods. Furthermore, our discrepancy in the utterance-level and recording-level results also align with results of [43, 44, 45], where it is hypothesised that there are too much random variation on the utterance-level to estimate severity reliably.

Our second research question concerned the question whether there are other approaches available that require less labelled training data while giving similar performance on the speech evaluation task. The naive listener experiment in Section 4.2.5 showed that the severity ratings of naive listeners have a very high correlation with the expert listener's severity ratings. These results imply that it is possible to reliably, and cost-efficiently scale up the annotation of oral cancer speech for the prototyping of data-driven automatic objective speech severity evaluation approaches. Please note though that although naive listeners are able to rate severity similarly to trained SLPs, other as-

[8] Note that this is not referring to the SynthNorm approaches, where we use synthetic speech in the evaluation process, but rather to the traditional process of TTS evaluation.

pects of speech, such as breathiness, nasality, hoarseness are likely not well rated by naive listeners but rather would require an expert listener. We also found that using binary labels indicating the presence or absence of oral cancer speech led to a reduced labelling effort but also nearly always resulted in a lower correlation with the human ratings than using the full 5-point scale ratings. For severity ratings, we thus advice to use a graded scale rather than binary labels.

Although our results are not yet well enough for clinical use, we think that our results do have a real-world applicability. For instance, these results are potentially good enough for use in smartphone applications: (1) The ASR and the LASSO models presented here have relatively low computational complexity compared to fully deep learning based methods such as [8], which is important due to the low memory requirements of smartphone devices. (2) In a smartphone use case, various noises and unexpected (conversational, spontaneous) speech modalities can be present. As our approaches have been tested with these scenarios, we are confident that performance will not deteriorate significantly in these conditions. Still, a more controlled test would be imminent, where similar speech recordings should be tested under different, real life, controlled noise conditions. For these smartphone scenarios, we suggest using the modulation spectrum regressor, if no reference transcription is available, and the Baseline+OC+ngram ASR model when a reference transcription is available. The Baseline+ngram+TTS method also includes a TTS pipeline inside it, which would likely cause additional difficulties when deploying to a device with a low memory requirement.

4.6. CONCLUSION

In this paper, we aimed to find the best method for the automatic evaluation of the severity of oral cancer speech. To do that, we collected a publicly available spontaneous oral cancer speech corpus. We compared two sets of reference-based methods and one set of reference-free methods. We evaluated our reference-free results on the level of entire recordings and on the level of single utterances. Our extensive experiments showed:

(1) two ASR models were found to have the highest correlation with the human ratings, when we have a single utterance and its transcription (reference-based): The Baseline+OC+ngram ASR model, an ASR model which uses oral cancer data during training and an n-gram based language model and the Baseline+ngram+TTS, which does not use oral cancer data during training, but uses synthetic speech references in the evaluation. When we use multiple utterances for rating without using a transcription (reference-free recording-level), a LASSO regression model was found to be the best using modulation spectrum features. (2) In an effort to reduce labelling effort, we found that naive listeners' ratings, e.g., obtained through crowd-sourcing, can be used instead of those of an expert listeners as their ratings were highly similar. Therefore, we encourage the usage of naive listener scores for speech severity labelling to reduce data collection costs, and therefore prototype automatic speech severity evaluation systems more efficiently.

4.7. ACKNOWLEDGEMENTS

We would like to thank Noa Hannah for helping out with the SLP ratings. This project has received funding from the European Union's Horizon 2020 research and innovation pro-

gramme under Marie Sklodowska-Curie grant agreement No 766287. The Department of Head and Neck Oncology and surgery of the Netherlands Cancer Institute receives a research grant from Atos Medical (Hörby, Sweden), which contributes to the existing infrastructure for quality of life research.

4.8. APPENDIX

Table 4.6: Wilcoxon signed-rank test between the scores given for the different recordings

	id001		id002						id003	id004				id005	id007	id008		id010	id011							id013	id014	id015	id016	id017	id018	id019
	1	3	8	12	14	16	19	25	10	11	15	27	29	18	21	23	24	31	4	5	6	7	13	22	28	17	30	32	33	34	35	36
1	same	***	***	***	***	***	***	***	***	***	***	***	***	***	***	***	***	***	***	***	***	***	***	***	***	***	***	***	***	***	***	***
3	***	same	***	***	***	***	***	***	***	***	***	***	***	***	0.32	***	***	0.02	***	***	***	0.56	0.18	***	***	***	***	***	***	***	***	***
8	***	***	same	***	***	0.01	0.06	***	***	***	0.32	0.05	0.08	***	***	***	***	***	***	***	***	***	***	***	***	***	***	***	***	***	***	***
12	***	***	***	same	***	1	0.32	***	***	0.03	0.01	0.1	0.03	***	***	***	***	***	***	***	***	***	***	***	***	***	***	***	***	***	***	***
14	***	***	***	***	same	***	***	***	***	***	***	***	***	***	***	***	***	***	***	***	***	***	***	***	***	***	***	***	***	***	***	***
16	***	***	0.01	1	***	same	0.08	***	***	0.13	0.02	0.05	0.06	***	***	***	***	***	***	***	***	***	***	***	***	***	***	***	***	***	***	***
19	***	***	0.06	0.32	***	0.08	same	***	***	0.03	0.1	0.32	0.32	***	***	***	***	***	***	***	***	***	***	***	***	***	***	***	***	***	***	***
25	***	***	***	***	***	***	***	same	***	***	***	***	***	***	***	***	***	***	***	***	***	***	***	***	***	***	***	***	***	***	***	***
10	***	***	***	***	***	***	***	same	***	***	***	***	***	***	***	***	***	***	***	***	***	***	***	***	***	***	***	***	***	***	***	***
11	***	***	***	0.03	***	0.13	0.03	***	***	same	***	0.01	***	***	***	***	***	***	***	***	***	***	***	***	***	0.01	***	***	***	***	***	***
15	***	***	0.32	0.01	***	0.02	0.1	***	***	***	same	0.08	0.16	***	***	***	***	***	***	***	***	***	***	***	***	***	***	***	***	***	***	***
27	***	***	0.05	0.1	***	0.05	0.32	***	***	0.01	0.08	same	0.32	***	***	***	***	***	***	***	***	***	***	***	***	***	***	***	***	***	***	***
29	***	***	0.08	0.03	***	0.06	0.32	***	***	0.16	0.32	same	***	***	***	***	***	***	***	***	***	***	***	***	***	***	***	***	***	***	***	***
18	***	***	***	***	***	***	***	***	***	***	***	***	same	***	***	***	***	***	***	***	***	***	***	***	***	***	***	***	***	***	***	***
21	***	0.32	***	***	***	***	***	***	***	***	***	***	***	same	***	***	***	0.03	***	***	***	1	0.06	***	***	***	***	***	***	***	***	***
23	***	***	***	***	***	***	***	***	***	***	***	***	***	***	***	same	***	***	***	***	***	***	***	***	***	***	***	***	***	***	0.01	***
24	***	***	***	***	***	***	***	***	***	***	***	***	***	***	***	***	same	***	***	***	***	***	***	***	***	***	***	***	***	***	***	***
31	***	0.02	***	***	***	***	***	***	***	***	***	***	***	***	0.03	***	***	same	***	***	***	0.06	***	***	***	***	***	***	***	***	***	***
4	***	***	***	***	***	***	***	***	***	***	***	***	***	***	***	***	***	***	same	0.16	***	***	***	***	***	***	***	***	***	***	***	***
5	***	***	***	***	***	***	***	***	***	***	***	***	***	***	***	***	***	***	0.16	same	***	***	***	***	***	***	***	***	***	***	***	***
6	***	***	***	***	***	***	***	***	***	***	***	***	***	***	***	***	***	***	***	***	same	***	***	***	***	***	***	***	***	***	***	***
7	***	0.56	***	***	***	***	***	***	***	***	***	***	***	***	1	***	***	0.06	***	***	***	same	0.13	***	***	***	***	***	***	***	***	***
13	***	0.18	***	***	***	***	***	***	***	***	***	***	***	***	0.06	***	***	***	***	***	***	0.13	same	***	***	***	***	***	***	***	***	***
22	***	***	***	***	***	***	***	***	***	***	***	***	***	***	***	***	***	***	***	***	***	***	***	same	***	***	***	***	***	***	***	***
28	***	***	***	***	***	***	***	***	***	***	***	***	***	***	***	***	***	***	***	***	***	***	***	***	same	***	***	***	***	***	***	***
17	***	***	***	***	***	***	***	***	***	0.01	***	***	***	***	***	***	***	***	***	***	***	***	***	***	***	same	***	***	***	***	***	***
30	***	***	***	***	***	***	***	***	***	***	***	***	***	***	***	***	***	***	***	***	***	***	***	***	***	***	same	***	***	***	***	***
32	***	***	***	***	***	***	***	***	***	***	***	***	***	***	***	***	***	***	***	***	***	***	***	***	***	***	***	same	***	***	***	***
33	***	***	***	***	***	***	***	***	***	***	***	***	***	***	***	***	***	***	***	***	***	***	***	***	***	***	***	***	same	***	***	0.56
34	***	***	***	***	***	***	***	***	***	***	***	***	***	***	***	***	***	***	***	***	***	***	***	***	***	***	***	***	***	same	***	***
35	***	***	***	***	***	***	***	***	***	***	***	***	***	***	***	0.01	***	***	***	***	***	***	***	***	***	***	***	***	***	***	same	***
36	***	***	***	***	***	***	***	***	***	***	***	***	***	***	***	***	***	***	***	***	***	***	***	***	***	***	***	***	0.56	***	***	same

BIBLIOGRAPHY

[1] K. D. Shield, J. Ferlay, A. Jemal, R. Sankaranarayanan, A. K. Chaturvedi, F. Bray, and I. Soerjomataram, "The global incidence of lip, oral cavity, and pharyngeal cancers by subsite in 2012," *CA: A Cancer Journal for Clinicians*, vol. 67, no. 1, pp. 51–64, 2017.

[2] O. C. Foundation, "Oral cancer facts," Feb 2019. [Online]. Available: https://oralcancerfoundation.org/facts/

[3] E. C. Ward and C. J. van As-Brooks, *Head and Neck Cancer: treatment, rehabilitation, and outcomes. Chapter 5: Speech and Swallowing Following Oral, Oropharyngeal, and Nasopharyngeal Cancer.* Plural Publishing, 2014.

[4] R. N. Rinkel, I. M. V.-d. Leeuw, E. J. van Reij, N. K. Aaronson, and C. R. Leemans, "Speech handicap index in patients with oral and pharyngeal cancer: better understanding of patients' complaints," *Head & Neck: Journal for the Sciences and Specialties of the Head and Neck*, vol. 30, no. 7, pp. 868–874, 2008.

[5] R. I. Zraick, G. B. Kempster, N. P. Connor, S. Thibeault, B. K. Klaben, Z. Bursac, C. R. Thrush, and L. E. Glaze, "Establishing validity of the consensus auditory-perceptual evaluation of voice (CAPE-V)," *American Journal of Speech-Language Pathology*, 2011.

[6] J. Oates, "Auditory-perceptual evaluation of disordered voice quality," *Folia Phoniatrica et Logopaedica*, vol. 61, no. 1, pp. 49–56, 2009.

[7] M. Windrich, A. Maier, R. Kohler, E. Nöth, E. Nkenke, U. Eysholdt, and M. Schuster, "Automatic quantification of speech intelligibility of adults with oral squamous cell carcinoma," *Folia Phoniatrica et Logopaedica*, vol. 60, no. 3, pp. 151–156, 2008.

[8] S. Quintas, J. Mauclair, V. Woisard, and J. Pinquier, "Automatic Prediction of Speech Intelligibility Based on X-Vectors in the Context of Head and Neck Cancer," in *Proc. Interspeech 2020*, 2020, pp. 4976–4980.

[9] V. Wolfe, R. Cornell, and J. Fitch, "Sentence/vowel correlation in the evaluation of dysphonia," *Journal of Voice*, vol. 9, no. 3, pp. 297–303, 1995.

[10] J. Revis, A. Giovanni, F. Wuyts, and J.-M. Triglia, "Comparison of different voice samples for perceptual analysis," *Folia Phoniatrica et Logopaedica*, vol. 51, no. 3, pp. 108–116, 1999.

[11] B. M. Halpern, R. van Son, M. van den Brekel, and O. Scharenborg, "Detecting and Analysing Spontaneous Oral Cancer Speech in the Wild," in *Proc. Interspeech 2020*, 2020, pp. 4826–4830. [Online]. Available: http://dx.doi.org/10.21437/Interspeech.2020-1598

[12] A. Tripathi, S. Bhosale, and S. K. Kopparapu, "A novel approach for intelligibility assessment in dysarthric subjects," in *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2020, pp. 6779–6783.

[13] B. M. Halpern, S. Feng, R. van Son, M. van den Brekel, and O. Scharenborg, "Low-resource automatic speech recognition and error analyses of oral cancer speech," *Speech Communication*, vol. 141, pp. 14–27, 2022.

[14] E. Vincent, R. Gribonval, and C. Févotte, "Performance measurement in blind audio source separation," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 14, no. 4, pp. 1462–1469, 2006.

[15] C. H. Taal, R. C. Hendriks, R. Heusdens, and J. Jensen, "A short-time objective intelligibility measure for time-frequency weighted noisy speech," in *2010 IEEE International Conference on Acoustics, Speech and Signal Processing*. IEEE, 2010, pp. 4214–4217.

[16] P. Janbakhshi, I. Kodrasi, and H. Bourlard, "Pathological speech intelligibility assessment based on the short-time objective intelligibility measure," in *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2019, pp. 6405–6409.

[17] R. Kubichek, "Mel-cepstral distance measure for objective speech quality assessment," in *Proceedings of IEEE Pacific Rim Conference on Communications Computers and Signal Processing*, vol. 1. IEEE, 1993, pp. 125–128.

[18] L. K. Smith and A. M. Goberman, "Long-time average spectrum in individuals with Parkinson disease," *NeuroRehabilitation*, 2014.

[19] S. Master, N. De Biase, V. Pedrosa, and B. M. Chiari, "The long-term average spectrum in research and in the clinical practice of speech therapists," *Pro-fono : revista de atualizacao cientifica*, 2006.

[20] K. Tanner, N. Roy, A. Ash, and E. H. Buder, "Spectral moments of the long-term average spectrum: Sensitive indices of voice change after therapy?" *Journal of Voice*, 2005.

[21] B. M. Halpern, J. Fritsch, E. Hermann, R. van Son, O. Scharenborg, and M. Magimai-Doss, "An objective evaluation framework for pathological speech synthesis," in *Speech Communication; 14th ITG Conference*. VDE, pp. 1–5.

[22] D. Martínez, P. D. Green, and H. Christensen, "Dysarthria intelligibility assessment in a factor analysis total variability space," in *Proc. Interspeech 2013*, 2013, pp. 2133–2137.

[23] I. Laaridh, W. B. Kheder, C. Fredouille, and C. Meunier, "Automatic Prediction of Speech Evaluation Metrics for Dysarthric Speech," in *Proc. Interspeech 2017*, 2017, pp. 1834–1838.

[24] L. Wan, Q. Wang, A. Papir, and I. L. Moreno, "Generalized end-to-end loss for speaker verification," in *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2018, pp. 4879–4883.

[25] T. Toda and K. Tokuda, "A speech parameter generation algorithm considering global variance for hmm-based speech synthesis," *IEICE TRANSACTIONS on Information and Systems*, vol. 90, no. 5, pp. 816–824, 2007.

[26] S. Takamichi, T. Toda, G. Neubig, S. Sakti, and S. Nakamura, "A postfilter to modify the modulation spectrum in hmm-based speech synthesis," in *2014 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2014, pp. 290–294.

[27] J. Kominek and A. W. Black, "The CMU Arctic speech databases," in *Proc. 5th ISCA Workshop on Speech Synthesis (SSW 5)*, 2004, pp. 223–224.

[28] F. Rudzicz, A. K. Namasivayam, and T. Wolff, "The torgo database of acoustic and articulatory speech from speakers with dysarthria," *Language Resources and Evaluation*, vol. 46, no. 4, pp. 523–541, 2012.

[29] R. Tibshirani, "Regression shrinkage and selection via the lasso," *Journal of the Royal Statistical Society: Series B (Methodological)*, vol. 58, no. 1, pp. 267–288, 1996.

[30] H. Bredin, R. Yin, J. M. Coria, G. Gelly, P. Korshunov, M. Lavechin, D. Fustes, H. Titeux, W. Bouaziz, and M.-P. Gill, "Pyannote. audio: neural building blocks for speaker diarization," in *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2020, pp. 7124–7128.

[31] J. M. Coria, H. Bredin, S. Ghannay, and S. Rosset, "A Comparison of Metric Learning Loss Functions for End-To-End Speaker Verification," in *Statistical Language and Speech Processing*, L. Espinosa-Anke, C. Martín-Vide, and I. Spasić, Eds. Springer International Publishing, 2020, pp. 137–148.

[32] D. Snyder, D. Garcia-Romero, G. Sell, D. Povey, and S. Khudanpur, "X-vectors: Robust dnn embeddings for speaker recognition," in *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2018, pp. 5329–5333.

[33] M. Ravanelli and Y. Bengio, "Speaker recognition from raw waveform with sincnet," in *2018 IEEE Spoken Language Technology Workshop (SLT)*. IEEE, 2018, pp. 1021–1028.

[34] R. Yamamoto, Y. Yamada, hyama5, Aria-K-Alethia, R. Huang, Hiroshiba, J. Regan, M. Roszkowski, and T. Shirani, "r9y9/nnmnkwii: v0.1.0 release," Aug. 2021. [Online]. Available: https://doi.org/10.5281/zenodo.5178769

[35] P. Janbakhshi, I. Kodrasi, and H. Bourlard, "Synthetic speech references for automatic pathological speech intelligibility assessment," in *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2020, pp. 6099–6103.

[36] D. B. Paul and J. M. Baker, "The design for the wall street journal-based CSR corpus," in *Proceedings of the workshop on Speech and Natural Language*. Association for Computational Linguistics, 1992, pp. 357–362.

[37] M. J. Gales, "Maximum likelihood linear transformations for hmm-based speech recognition," *Computer speech & language*, vol. 12, no. 2, pp. 75–98, 1998.

[38] J. Shen, R. Pang, R. J. Weiss, M. Schuster, N. Jaitly, Z. Yang, Z. Chen, Y. Zhang, Y. Wang, R. Skerrv-Ryan *et al.*, "Natural tts synthesis by conditioning wavenet on mel spectrogram predictions," in *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2018, pp. 4779–4783.

[39] B. McFee, V. Lostanlen, A. Metsai, M. McVicar, S. Balke, C. Thomé, C. Raffel, F. Zalkow, A. Malek, Dana, K. Lee, O. Nieto, J. Mason, D. Ellis, E. Battenberg, S. Seyfarth, R. Yamamoto, K. Choi, viktorandreevichmorozov, J. Moore, R. Bittner, S. Hidaka, Z. Wei, nullmightybofo, D. Hereñú, F.-R. Stöter, P. Friesch, A. Weiss, M. Vollrath, and T. Kim, "librosa/librosa: 0.8.0," Jul. 2020. [Online]. Available: https://doi.org/10.5281/zenodo.3955228

[40] M. Mashimo, T. Toda, K. Shikano, and N. Campbell, "Evaluation of cross-language voice conversion based on GMM and straight," in *Proc. 7th European Conference on Speech Communication and Technology (Eurospeech 2001)*, 2001, pp. 361–364.

[41] H. Kim and S. Nanney, "Familiarization effects on word intelligibility in dysarthric speech," *Folia Phoniatrica et Logopaedica*, vol. 66, no. 6, pp. 258–264, 2014.

[42] H. Kim, "Familiarization effects on consonant intelligibility in dysarthric speech," *Folia Phoniatrica et Logopaedica*, vol. 67, no. 5, pp. 245–252, 2015.

[43] R. Clapham, C. Middag, F. Hilgers, J.-P. Martens, M. Van Den Brekel, and R. Van Son, "Developing automatic articulation, phonation and accent assessment techniques for speakers treated for advanced head and neck cancer," *Speech Communication*, vol. 59, pp. 44–54, 2014.

[44] R. P. Clapham, J.-P. Martens, R. J. van Son, F. J. Hilgers, M. M. van den Brekel, and C. Middag, "Computing scores of voice quality and speech intelligibility in tracheoesophageal speech for speech stimuli of varying lengths," *Computer Speech & Language*, vol. 37, pp. 1–10, 2016.

[45] C. Middag, J.-P. Martens, G. Van Nuffelen, and M. De Bodt, "Automated intelligibility assessment of pathological speech using phonological features," *EURASIP Journal on Advances in Signal Processing*, vol. 2009, pp. 1–9, 2009.

5

PATHOLOGICAL VOICE ADAPTATION WITH AUTOENCODER-BASED VOICE CONVERSION

In this paper, we propose a new approach to pathological speech synthesis. Instead of using healthy speech as a source, we customise an existing pathological speech sample to a new speaker's voice characteristics. This approach alleviates the evaluation problem one normally has when converting typical speech to pathological speech, as in our approach, the voice conversion (VC) model does not need to be optimised for speech degradation but only for the speaker change. This change in the optimisation ensures that any degradation found in naturalness is due to the conversion process and not due to the model exaggerating characteristics of a speech pathology. To show a proof of concept of this method, we convert dysarthric speech using the UASpeech database and an autoencoder-based VC technique. Subjective evaluation results show reasonable naturalness for high intelligibility dysarthric speakers, though lower intelligibility seems to introduce a marginal degradation in naturalness scores for mid and low intelligibility speakers compared to ground truth. Conversion of speaker characteristics for low and high intelligibility speakers is successful, but not for mid. Whether the differences in the results for the different intelligibility levels is due to the intelligibility levels or due to the speakers needs to be further investigated.

5.1. INTRODUCTION

Data-driven speech synthesis has recently been reaching new heights with the introduction of deep neural networks (DNNs). However, the success of these techniques is sub-

This chapter has been published as: Illa, M., Halpern, B. M., van Son, R., Moro-Velázquez, L., & Scharenborg, O. (2021). Pathological voice adaptation with autoencoder-based voice conversion. In 11th ISCA Speech Synthesis Workshop (pp. 19-24). ISCA. The PhD candidate contributed to the supervision, experiment design, and the evaluation of the project.

ject to high quality data and a large quantity of data, either of which is not available for many applications. Pathological speech synthesis, where the goal is to synthesise natural, but pathologically sounding samples, is such an application. Pathological speech synthesis has several motivations, the most notable being the data augmentation for automatic speech recognisers (ASRs), where the goal is to generate more data in order to improve recognition of pathological speech [1, 2, 3]. The second motivation for the development of pathological speech synthesis is that it could assist in informed decision making for the medical conditions at the root of the pathology. For instance, oral cancer surgery results in changes to a speaker's voice. The availability of a synthesis model that can generate how the voice could sound after surgery could help the patients and clinicians to make informed decisions about the surgery and alleviate the stress of the patients [4, 5].

While there are many speech synthesis techniques for typical speech, not many of these are applicable if we wish to synthesise highly natural pathological speech. Formant [6] and articulatory synthesis [7] are lacking in naturalness compared to DNN-based speech synthesis. Text-to-speech techniques (TTS) lack both linguistic resources (i.e a pronunciation lexicon) and the amount of data needed for these problems. The only promising method to synthesise pathological speech seems to be voice conversion (VC), which only needs a relatively small amount of data, compared to neural TTS.

However, synthesising pathological speech via VC is not without challenges. Existing pathological speech corpora [8, 9, 5, 10] provide healthy control speakers, but healthy speech recordings from the same pathological speaker are rarely available. This means that a successful pathological voice conversion system needs to learn conversion of both, the voice and pathological characteristics simultaneously, as suggested in previous work [4]. However, evaluation of such a setup is difficult. This is because the VC system is directly optimised for speech degradation in terms of the pathology, which would need the listeners (the evaluators of these systems) to be able to rate the success of generating the pathological characteristics and the synthetic/natural aspects of the speech separately. As we will show later in this paper, listeners struggle differentiating between speech severity and synthetic aspects of the speech. This can result in two, counter-intuitive scenarios from the viewpoint of typical VC: (1) a pathological VC system that is not able to properly capture the characteristics of the pathological speech could still receive better naturalness scores than the reference pathological speech; (2) Conversely, a VC system that is able to mimic the pathology, albeit exaggeratedly, could produce a naturalness score that is a lot lower than that of the reference.

Therefore, we propose a new approach where instead of using healthy speech as source for the VC, we use dysarthric speech, which is already pathological, and the VC system only has to customise it to a new (healthy/dysarthric) speaker's voice characteristics, i.e by using some representation of the speaker (speaker embedding). This synthesis approach alleviates the problem with naturalness ratings as the dysarthric-to-dysarthric VC is not optimised directly for speech degradation, therefore degradation is only due to the synthetic aspects compared to the source pathological utterance. Our first goal is to assess whether we can convert the voice characteristics of the pathological speakers in this setup in a natural way, while simultaneously assessing how natural real pathological speech is perceived.

In order to perform the VC, an autoencoder-based method will be used [11]. Autoencoder-based methods are of special interest in clinical scenarios as they are non-parallel, thus allow for incomplete data collection situations, while also being easier to train than GAN-based methods due to well-defined convergence criteria because they have only a single loss [12, 13, 14]. In this paper, we use HL-VQ-VAE-3 which is a type of variational autoencoder (VAE) using discrete representations. This hierarchical design has recently shown to give better results for VC [15] than the original VQ-VAE. Furthermore, by conditioning on speaker labels, the model allows converting to/from multiple speakers within one single model.

An important additional goal of this work is to investigate whether standard VC techniques can be used for non-standard speech. It is well known from other domains of speech technology such as automatic speech recognition (ASR) that standard ASR systems perform poorly on atypical speech [16, 17, 18, 19, 20], making standard speech technology techniques less accessible to people with atypical speech. Our paper is thus also a preliminary investigation of a VQ-VAE-based VC technique's performance on converting a pathological source utterance instead of a typical utterance from a non-dysarthric speaker.

To summarise, in this paper we train a dysarthric-to-dysarthric VC system to answer the following research questions: (**RQ1**) *Can we convert the voice characteristics of a pathological speaker to another pathological speaker of the same severity with reasonable naturalness (where reasonable means comparable to non-parallel VC methods on typical speech)? In other words, is VC technology accessible to people with pathological speech?* (**RQ2**) *How does (real) pathological speech affect the mean opinion score (MOS)? In other words, what is the maximum attainable naturalness of synthetic pathological speech?*

Section 5.2 will start with the discussion of the used UASpeech dataset and the used VQ-VAE methods for the task, and finally concluded by the experimental design to test the approach. The perceptual evaluation results are presented in Section 5.3, followed by a discussion of the limitations of the proposed method, and further comments on the accessibility of VC to pathological speakers. Some of the samples are available at *https://pathologicalvc.github.io* .

5.2. DESIGN AND METHODS

5.2.1. DESCRIPTION OF THE DATASET AND PREPROCESSING

In this study we use the UASpeech corpus [8], which contains isolated-word recordings of 15 speakers with dysarthria. These recordings consist of 449 words which are divided into 3 blocks of equal length (B1, B2 and B3). The speakers are divided into four groups based on their intelligibility: very low, low, mid and high, which correspond to 0-25%, 25-50%, 50-75% and 75-100% human transcription word error rate (WER) of the recordings, respectively. The transcriptions were done by 5 American English native speakers, who are non-expert listeners.

The vocoder used (see Section 5.2.2) is trained using the VCTK dataset [21], which contains speech of 108 native English speakers with different accents. The preprocessing consists of downsampling the tracks from 48 kHz to 24 kHz, which is done with librosa [22].

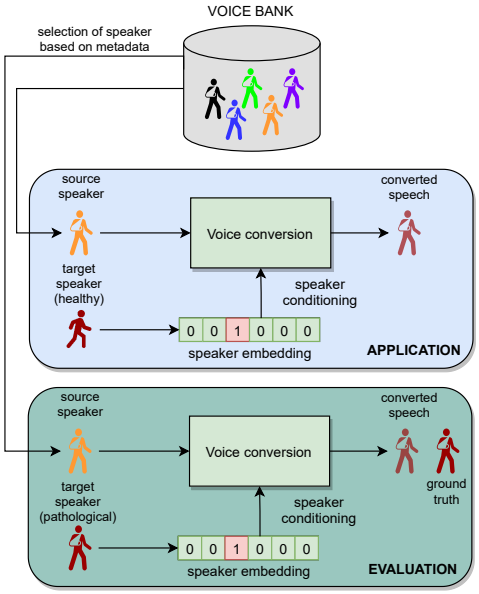


Figure 5.1: Outline of our approach: the speech from a model pathological speaker is converted into speech with the characteristics of another pathological speaker. Red/orange colours denote the identity of the speaker. The figure is further explained in Section 5.2.

The UASpeech data is preprocessed following [2]: stationary noise is removed using Noisereduce [23] and the silence from the beginning and end of the clips is cut. Then, the audio is resampled from 16 kHz to 24 kHz and normalised. Finally, 80-dimension mel-spectrograms (similar to [24]) are extracted from the audio files and used to compute the mel-cepstrum, which serves as input to our model.

5.2.2. VOICE CONVERSION MODEL

The model is a 3-stage VQ-VAE. In the first stage, the input x to the model is a mel-cepstrum that goes through the convolutional encoder resulting in a hidden variable u_1 and a latent variable z_1. The second stage is identical to the first stage, except instead of x, now u_1 is fed into another convolutional encoder, resulting in u_2 and z_2. This is repeated for the third stage, feeding u_2 to obtain z_3 and u_3. This successive encoding serves to model the features in the speech that are present on successively longer temporal scales.

The variables z_n are all quantised using a nearest neighbour classifier with respect to the codebook's codewords of the corresponding stage. Then, we perform the decoding of the quantised variables q_n at each stage. The decoder is also convolutional which is additionally conditioned on a speaker label. During training, a speaker embedding table is learned from the training speakers, and during conversion/inference, this embedding will correspond to the target speaker of the conversion, which we can get by a table lookup. The decoding starts at the third stage and goes back to the first stage. The

input of the third stage decoder is q_3 while for the second and first level the q_n signal is concatenated with the output v_n of the previous stage (the output v_2 of the 3rd stage is fed to the 2nd and the output v_1 of the 2nd is fed to the 1st).

For the conversion, the trained model receives the input mel-cepstrum from a source speaker which is encoded and quantised in the same way as it is during training. Then, the speaker embedding is used to condition the decoder on a target speaker, so the source speaker quantised latent variables q_n are decoded conditioned on the target speaker embedding, which results in the decoded mel-cepstrum. Finally, the mel-cepstrum is resynthesised to the speech waveform using a Parallel WaveGAN vocoder[1] [25].

5.2.3. DETAILS OF THE EXPERIMENTAL DESIGN

As a reminder, in this study, we customise pathological speech to a different pathological speaker's voice characteristics. However, the clinical application would need customisation to a healthy speaker's characteristics. In the top panel of Figure 5.1, the application scenario is visualised, i.e., how the system could be used in a clinical setting. In the bottom part, our proposed evaluation scenario - the experiments that we do in the paper - is illustrated.

Looking at the top panel, a source pathological speaker is first selected from a large voice bank consisting of many samples of pathological speakers. Based on metadata, a clinical team could decide the kind of pathological speech degradation which is most likely for a patient. In this paper, we pair up by severity, but in actual practice an appropriate source speaker could be found matched by age, region, and type of treatment. This leads to a selection of a source pathological speaker. Using a small amount of a new patient's voice (target speaker), a speaker embedding can be extracted using the VQ-VAE based technique. Finally, we obtain the converted speech, which is expected to be pathological, but with the new patient's voice characteristics. The problem is that for the UASpeech, we don't have parallel pre-pathology and post-pathology voices. Therefore, a separate evaluation scheme has to be setup where we assume that the pathological and the healthy speaker embeddings should be unchanged for the same speaker, which is not always true, we refer to further discussion about this in Section 5.3.3.

The evaluation scheme is explained in the bottom panel. To circumvent the problem with the pre-pathology and post-pathology, we change the conversion process for the evaluation as follows. Instead of a new healthy speaker, we enrol a new dysarthric speaker with a matched intelligibility of the speech pathology from the UASpeech dataset because a ground truth (GT) is available there. The converted speech can then be compared to this GT to provide a proof of concept for the system.

Table 5.1: Speaker pairs used for the VC experiments and their subjective WER differences.

Speaker A (WER%)	**Speaker B (WER%)**	**ΔWER (%)**
M04 (2%)	M12 (7.4%)	5.4%
M05 (58%)	M11 (62%)	4%
M08 (93%)	M10 (93%)	0%

[1] https://github.com/kan-bayashi/ParallelWaveGAN

In our experiments, we convert the speech of three speaker pairs in both directions The setup for the experiments is the following. We train the VC model with all B1 and B3 sets of words of every dysarthric speaker to stay consistent with the standard UASpeech train-test partitioning.

We perform VC on the speech from B2 between speakers with a similar level of dysarthria. The selected dysarthric speaker pairs along with their corresponding human transcription error rates from UASpeech are summarised in Table 5.1. Unfortunately, it has not been possible to include females speakers because all female speakers had a different severity in the UASpeech dataset. We also refrained from controlling for the type of dysarthria in our experimental design, as that would have led to certain speaker pairs having excessive difference in their intelligibility, which would contrive the aim of the paper.

5.2.4. SUBJECTIVE EVALUATION EXPERIMENTS

In order to answer our research questions, we performed subjective evaluation experiments. For RQ1 a subjective speaker similarity experiment was carried out, while for RQ2 a subjective naturalness experiment was carried out. The design of these experiments (including the composition of different stimuli) closely follow those of the VCC challenge standards [26, 27]. These experiments were run on the Qualtrics platform, and the participants (10 American English native listeners) were recruited through Prolific. All participants were remunerated justly (7.80 GBP per hour).

For the naturalness experiment, we used a mean opinion score (MOS) naturalness test. We hypothesised that listeners will not be able to distinguish between the distortions in the audio and the pathological characteristics of the speech. In order to account for this, we included GT stimuli in the naturalness test, which allows direct comparison of naturalness with real samples. The GT shows the maximum attainable naturalness (second part of RQ2) and the differences of the GT and VC scores show the reduction due to the synthetic aspects. To answer the first part of RQ2, we included healthy, natural stimuli, which allows us to measure the reduction in naturalness due to the reduction intelligibility. Nevertheless, we encouraged listeners to ignore the atypical aspects of the speech by adopting the naturalness question from the VCC2020 [26], which was proposed for cross-lingual VC, where pronunciation errors could appear, similar to pathological speech. For the speaker similarity test, we used an AB test in which listeners were asked to listen to two stimuli, and indicate if they thought they came from the same speaker, and rate their confidence in this decision. The question for the speaker similarity was directly adopted from the VCC2016 challenge [27].

5.3. RESULTS AND DISCUSSION

5.3.1. NATURALNESS

The results of the naturalness experiments are presented in Figure 5.2, which shows the MOS score for each of the seven types of speech tested, grouped by intelligibility, and with their 95% confidence intervals indicated. For clarity, the actual MOS scores are indicated on top of each bar.

We first focus on the question how GT pathological speech affects the naturalness

perceived by listeners which is measured by the MOS score (our RQ2). Figure 5.2 shows that healthy speech and GT high intelligibility dysarthric speech have a similar MOS score. However, as intelligibility decreases, so does the MOS score, indicating that the MOS score not only captures naturalness but is influenced by the intelligibility of the speech. These results show that naive listeners cannot separate severity of a pathology and unnaturalness when asked to judge the naturalness of a speech sample. This also means that the GT MOS results are an upper bound on the achievable naturalness of synthetic pathological samples.

Regarding the synthetic pathological speech, the performance on the high (VC) samples is somewhat lower than the performance of the HL-VQ-VAE-3 model on the VCC2020 challenge and identical to the performance of autoencoder-based models (2.1) [15]. However, the type of stimuli is different, so the differences in MOS are not directly comparable. The difference is most likely due to channel differences, the decreased intelligibility of the speech, and the different sampling frequency (UASpeech is 16 kHz, while VCC2020 is 24 kHz). When we compare the MOS scores for the converted speech of the different intelligibility speakers, we observe a slight degradation in naturalness with decreasing intelligibility. Comparing the VC and GT results, however, we observe a large degradation for the converted high intelligibility speech (Wilcoxon signed-rank test: $p \leq 0.05$). The difference in VC and GT MOS scores for the mid and low intelligibility speakers is much smaller (Wilcoxon signed-rank test: mid $p \leq 0.05$, low $p \geq 0.05$). It is possible that the standard 5-point MOS does not allow to express the nuances between mid and low samples appropriately. Therefore, for future studies concerning naturalness of pathological speech, we would recommend using a slightly wider, 7-point scale. Returning to RQ1, we can conclude that the synthetic speech of mid and low intelligibility pathological speakers have a naturalness that is perceived similar to that of real pathological speech, while synthetic high intelligibility pathological speech is not perceived as being as natural as real high intelligibility pathological speech.

To summarise, pathological speech is not perceived natural according to the MOS scale by naive listeners. In the case of mid and low intelligibility pathological speech, the perceived naturalness is similar between that of synthetic and real pathological speech. This is, however, not the case for high intelligibility synthesised pathological speech which is rated as being far less natural than real pathological speech. The performance of the VC approach is comparable to the one observed with typical speakers, therefore the current method is accessible to typical speakers, however this does not mean that VC is accessible to typical speakers (see Section 5.3.4).

5.3.2. SIMILARITY

This section presents and discusses the results of the similarity experiments in order to answer the question whether it is possible to convert voice characteristics of pathological speakers. The results are presented in Figure 5.3. In each of the 12 panels, we visualise the results of comparing a voice converted (VC-D / VC-S) sample with the GT source (S) (Similarity to source) or the GT target (Similarity to target). Also, the GT samples are compared between them: S samples are compared to S samples to know how recognisable the source speaker is, T samples are compared to T samples to know how recognisable the target speaker is and S samples are compared to T samples in order to know how

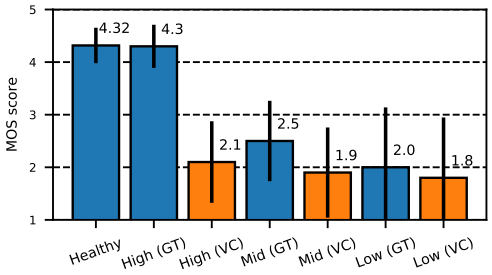


Figure 5.2: Mean opinion scores for naturalness grouped by intelligibility with 95% confidence intervals. Blue denotes original, while orange denotes VC samples.

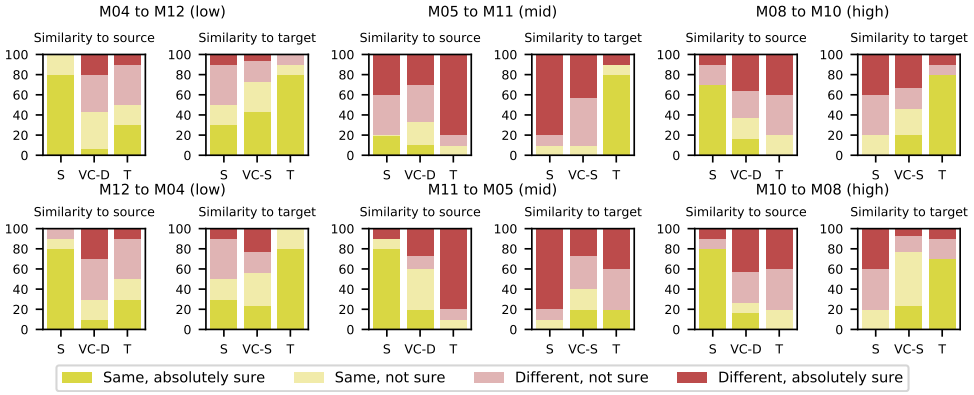


Figure 5.3: Results of the speaker similarity experiments grouped by intelligibility pairs. S stands for source, T for target, VC-D for voice conversion different (VC samples should be different from source) and VC-S for voice conversion same (VC samples should be same as target).

distinguishable is the source from the target speaker. Note that for each speaker pair in the top panel the source speaker is the target one in the bottom panel and vice versa, so this information appears repeated in Figure 5.3. Additionally letters in the case of the VC comparisons are used to help interpretation of the figures: VC-D stands for VC-different (i.e when converting M04 to M12, the converted should be different from M04), VC-S stands for VC-same (similarly, when converting M04 to M12, the converted should be same as M12).

For the low intelligibility pair (left 2 columns of Figure 3), the speakers seem reasonably distinguishable when looking at the GT as there is a 100% of agreement that M04 samples are produced by M04 and 90% for M12. For the speech samples of speaker M04 converted to speaker M12 (top panels), 73.33% of the converted samples were indicated as being from speaker M12 (VC-S), meaning that the conversion is fairly successful for that pair. For the speech samples of speaker M12 converted to speaker M04 (bottom panels), 56.33% of the converted M12-M04 samples (VC-S) were indicated as being from speaker M04. The results show that for the M12-M04 conversion the model is able to remove some of the source speaker (M12) characteristics and add some of the target

(M04) ones, although to a lesser extent than in the M04-M12 conversion. Therefore, we conclude that the voice characteristic conversions for the low intelligibility speakers are successful.

For the mid intelligibility pair (middle four panels), the M11 seems to be clearly recognisable as there is a 90% of agreement that M11 samples are produced by M11, however listeners have difficulties recognising the voice characteristics of M05, i.e., only 20% of the trials where both samples were from speaker M05 were judged as both being from M05. For M05-M11 the VC performs poorly, which is indicated by 90% perceiving it different from the target (VC-S result). For M05-M11 the VC-S reaches a 20% of absolutely sure agreement. Notice that although it is a low score, it is the same that the GT samples exhibit. The voice characteristic conversions for the mid intelligibility speakers are thus inconclusive: while in one case the VC fails, in the other participants fail to recognise the speaker even from the GT samples. Further experimentation with more speaker pairs is needed.

For the high intelligibility pairs (right 2 columns of Figure 3), the speakers seem reasonably distinguishable. We can see that there is a 70% of agreement that M08 samples are produced by M08 and an 80% for M10. For M08-M10, there is a 46.66% of agreement that the converted samples sound like M10. For M10 to M08 VC, 75% of the listeners indicate that the converted samples sound like M08. We can see that some of the voice characteristics are successfully transferred for the high intelligibility samples, however while on the conversions M10 to M08 the result is similar to the GT samples, on the other direction (M08 to M10) there is a gap of 33.33% with respect to the GT. This behaviour is the same that we observed with low intelligibility pair conversions: although the speakers from the same pair are recognised with a similar agreement (100% and 90% for low intelligibility and 80% and 70% for the high intelligibility) the conversions are more successful in one direction than on the other.

5.3.3. LIMITATIONS OF THE PROPOSED APPROACH

An assumption of the proposed approach is that the speaker identity is not affected by the speech pathology, which is certainly untrue for speech pathologies which are dysphonic, i.e. where the voice characteristics are known to be affected. By performing AB testing with GT speakers, we have tried to account for these scenarios in the perceptual evaluations. From the speaker similarity experiment, we have seen that in some cases (i.e., M05) listeners had difficulties of recognising the voice characteristics even in the GT. These results confirm that the proposed approach cannot be used for all types of speech pathologies. To solve this issue, we would need to have a deeper understanding of what happens to the speaker characteristics in these speech pathologies. For example, the speaker embeddings themselves could be used to predict the new pathological speaker embeddings of the same speaker, transformed according to the vocal pathology (i.e. type of dysphonia).

5.3.4. ACCESSIBILITY OF VC TO ATYPICAL SPEAKERS

VC of atypical speech produced similar naturalness in the high intelligibility case as typical speech on VQ-VAE based methods. Nevertheless, we see that there is room for improvement compared to typical speech, as other studies employing certain non-parallel

VC approaches can achieve human-like naturalness. Unfortunately, these VC approaches cannot easily be used for our task as they often leverage linguistic features or ASR bottleneck features [28, 29]. The need for ASR features is especially problematic as these features are extracted from ASR systems, whose performance on atypical speech is generally much worse than that on typical speech, meaning that the quality of these extracted features are also expected to be lower for these speakers. Therefore, we conclude that accessibility to VC is limited for atypical speakers, but this is because parallel and ASR-based techniques can hardly be used by them.

5.4. CONCLUSIONS

In this paper, we propose a new approach to pathological speech synthesis, by customising an existing pathological speech sample to a new speaker's voice characteristics. In order to do this pathological-to-pathological speech conversion, we use an autoencoder-based voice conversion (VC) technique. When comparing our results with the ones obtained in the VCC2020 challenge dataset [15], we can see that ours are somewhat lower, which is most likely due to channel differences, the decrease in the speech intelligibility and the different sampling rate. We find that even real pathological speech seems to affect perceived naturalness as shown by MOS scores, meaning that there is a bound on achievable naturalness for pathological speech conversion. Overall, we observe a decreasing trend in MOS with decreasing intelligibility. Therefore, for low and mid intelligibility, the difference in perceived naturalness between real and VC is small. Conversion of voice characteristics for low intelligibility speakers is successful, for high intelligibility it is also possible to transfer the voice characteristics partially. However, more experimentation is needed for the mid intelligibility with more speakers: we experienced that in one case the VC failed, and on the other participants fail to recognise the speaker even from the real recordings. Whether the differences in the results for the different intelligibility levels is due to the intelligibility levels or due to other speech characteristics needs to be further investigated. The question of pathological intergender (male to female) and female VC also needs to be investigated. The performance of the approach is comparable to the one observed with typical speakers, therefore the current method is accessible to atypical speakers. However, in the paper, we outlined some issues such as the need for linguistic resources and parallel data, as an obstacle for more natural VC for pathological speakers.

BIBLIOGRAPHY

[1] B. Vachhani, C. Bhat, and S. K. Kopparapu, "Data augmentation using healthy speech for dysarthric speech recognition." in *Interspeech*, 2018, pp. 471–475.

[2] J. Harvill, D. Issa, M. Hasegawa-Johnson, and C. Yoo, "Synthesis of new words for improved dysarthric speech recognition on an expanded vocabulary," in *International Conference on Acoustics, Speech and Signal Processing*, 2021.

[3] Y. Jiao, M. Tu, V. Berisha, and J. Liss, "Simulating dysarthric speech for training data augmentation in clinical speech applications," in *2018 IEEE international confer-*

ence on acoustics, speech and signal processing (ICASSP). IEEE, 2018, pp. 6009–6013.

[4] B. M. Halpern, J. Fritsch, E. Hermann, R. van Son, O. Scharenborg, and M. Magimai-Doss, "An objective evaluation framework for pathological speech synthesis," in *Speech Communication; 14th ITG Conference*. VDE, pp. 1–5.

[5] B. M. Halpern, R. van Son, M. van den Brekel, and O. Scharenborg, "Detecting and Analysing Spontaneous Oral Cancer Speech in the Wild," in *Proc. Interspeech 2020*, 2020, pp. 4826–4830. [Online]. Available: http://dx.doi.org/10.21437/Interspeech.2020-1598

[6] F. Rudzicz, "Adjusting dysarthric speech signals to be more intelligible," *Computer Speech & Language*, vol. 27, no. 6, pp. 1163–1177, 2013.

[7] S. Aryal and R. Gutierrez-Osuna, "Data driven articulatory synthesis with deep neural networks," *Computer Speech & Language*, vol. 36, pp. 260–273, 2016.

[8] H. Kim, M. Hasegawa-Johnson, A. Perlman, J. Gunderson, T. S. Huang, K. Watkin, and S. Frame, "Dysarthric speech database for universal access research," in *Ninth Annual Conference of the International Speech Communication Association*, 2008.

[9] F. Rudzicz, A. K. Namasivayam, and T. Wolff, "The TORGO database of acoustic and articulatory speech from speakers with dysarthria," *Language Resources and Evaluation*, vol. 46, no. 4, pp. 523–541, 2012.

[10] C. Middag, "Automatic analysis of pathological speech," Ph.D. dissertation, Ghent University, 2012.

[11] W.-N. Hsu, Y. Zhang, and J. Glass, "Learning latent representations for speech generation and transformation," in *Proc. Interspeech 2017*, 2017, pp. 1273–1277. [Online]. Available: http://dx.doi.org/10.21437/Interspeech.2017-349

[12] T. Kaneko and H. Kameoka, "Cyclegan-vc: Non-parallel voice conversion using cycle-consistent adversarial networks," in *2018 26th European Signal Processing Conference (EUSIPCO)*. IEEE, 2018, pp. 2100–2104.

[13] T. Kaneko, H. Kameoka, K. Tanaka, and N. Hojo, "CycleGAN-VC2: Improved CycleGAN-based non-parallel voice conversion," in *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2019, pp. 6820–6824.

[14] Kaneko, Takuhiro and Kameoka, Hirokazu and Tanaka, Kou and Hojo, Nobukatsu, "CycleGAN-VC3: Examining and Improving CycleGAN-VCs for Mel-Spectrogram Conversion," in *Proc. Interspeech 2020*, 2020, pp. 2017–2021. [Online]. Available: http://dx.doi.org/10.21437/Interspeech.2020-2280

[15] T. V. Ho and M. Akagi, "Non-parallel voice conversion based on hierarchical latent embedding vector quantized variational autoencoder," in *Proc. Joint Workshop for the Blizzard Challenge and Voice Conversion Challenge 2020*, 2020, pp. 140–144.

[16] S. Feng, O. Kudina, B. M. Halpern, and O. Scharenborg, "Quantifying bias in automatic speech recognition," *arXiv preprint arXiv:2103.15122*, 2021.

[17] M. Adda-Decker and L. Lamel, "Do speech recognizers prefer female speakers?" in *Ninth European Conference on Speech Communication and Technology*, 2005.

[18] L. Moro-Velazquez, J. Cho, S. Watanabe, M. A. Hasegawa-Johnson, O. Scharenborg, H. Kim, and N. Dehak, "Study of the Performance of Automatic Speech Recognition Systems in Speakers with Parkinson's Disease," in *Proc. Interspeech 2019*, 2019, pp. 3875–3879. [Online]. Available: http://dx.doi.org/10.21437/Interspeech.2019-2993

[19] A. Koenecke, A. Nam, E. Lake, J. Nudell, M. Quartey, Z. Mengesha, C. Toups, J. R. Rickford, D. Jurafsky, and S. Goel, "Racial disparities in automated speech recognition," *Proceedings of the National Academy of Sciences*, vol. 117, no. 14, pp. 7684–7689, 2020.

[20] E. Hermann and M. M. Doss, "Dysarthric speech recognition with lattice-free MMI," in *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2020, pp. 6109–6113.

[21] C. Veaux, J. Yamagishi, K. MacDonald *et al.*, "CSTR VCTK Corpus: English Multi-speaker Corpus for CSTR Voice Cloning Toolkit (2017)," *URL http://dx. doi. org/10.7488/ds*, 2017.

[22] B. McFee, C. Raffel, D. Liang, D. P. Ellis, M. McVicar, E. Battenberg, and O. Nieto, "librosa: Audio and music signal analysis in python," 2015.

[23] T. Sainburg, "timsainb/noisereduce: v1.0," Jun. 2019. [Online]. Available: https://doi.org/10.5281/zenodo.3243139

[24] R. Skerry-Ryan, E. Battenberg, Y. Xiao, Y. Wang, D. Stanton, J. Shor, R. Weiss, R. Clark, and R. A. Saurous, "Towards end-to-end prosody transfer for expressive speech synthesis with Tacotron," in *International Conference on Machine Learning*. PMLR, 2018, pp. 4693–4702.

[25] R. Yamamoto, E. Song, and J.-M. Kim, "Parallel WaveGAN: A fast waveform generation model based on generative adversarial networks with multi-resolution spectrogram," in *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2020, pp. 6199–6203.

[26] Z. Yi, W.-C. Huang, X. Tian, J. Yamagishi, R. K. Das, T. Kinnunen, Z. Ling, and T. Toda, "Voice conversion challenge 2020—intra-lingual semi-parallel and cross-lingual voice conversion—," in *Proc. Joint Workshop for the Blizzard Challenge and Voice Conversion Challenge 2020*, 2020, pp. 80–98.

[27] T. Toda, L.-H. Chen, D. Saito, F. Villavicencio, M. Wester, Z. Wu, and J. Yamagishi, "The voice conversion challenge 2016." in *Interspeech*, 2016, pp. 1632–1636.

[28] L.-J. Liu, Z.-H. Ling, Y. Jiang, M. Zhou, and L.-R. Dai, "WaveNet Vocoder with Limited Training Data for Voice Conversion," in *Proc. Interspeech 2018*, 2018, pp. 1983–1987. [Online]. Available: http://dx.doi.org/10.21437/Interspeech.2018-1190

[29] X. Tian, J. Wang, H. Xu, E. S. Chng, and H. Li, "Average Modeling Approach to Voice Conversion with Non-Parallel Data." in *Odyssey*, vol. 2018, 2018, pp. 227–232.

6

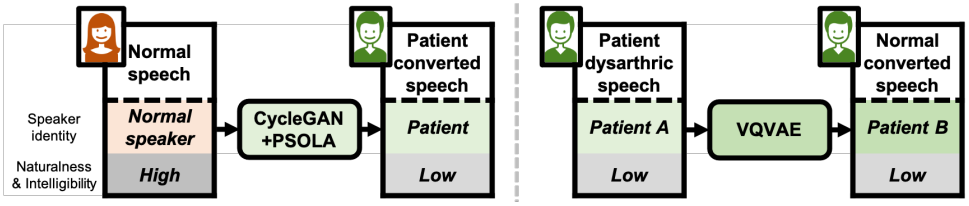
TOWARDS IDENTITY PRESERVING NORMAL TO DYSARTHRIC VOICE CONVERSION

We present a voice conversion framework that converts normal speech into dysarthric speech while preserving the speaker identity. Such a framework is essential for (1) clinical decision making processes and alleviation of patient stress, (2) data augmentation for dysarthric speech recognition. This is an especially challenging task since the converted samples should capture the severity of dysarthric speech while being highly natural and possessing the speaker identity of the normal speaker. To this end, we adopted a two-stage framework, which consists of a sequence-to-sequence model and a nonparallel frame-wise model. Objective and subjective evaluations were conducted on the UASpeech dataset, and results showed that the method was able to yield reasonable naturalness and capture severity aspects of the pathological speech. On the other hand, the similarity to the normal source speaker's voice was limited and requires further improvements.

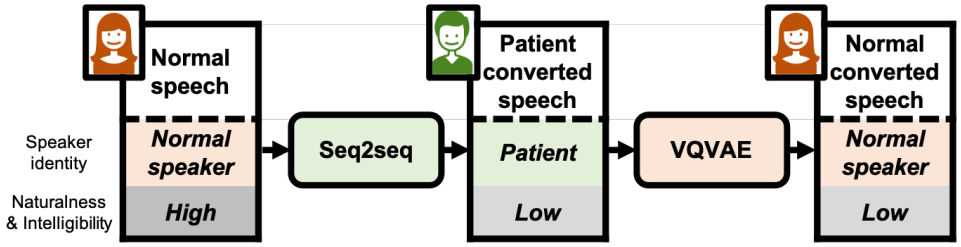
6.1. INTRODUCTION

Neural voice conversion (VC) has substantially improved the naturalness of synthesised speech in a wide range of tasks, including read speech [1], emotional speech [2] and whispered speech [3]. However, pathological VC (and TTS too) is a largely unexplored area, which has several interesting applications. In this work, we focus on normal-to-dysarthric (N2D) VC, which refers to the task of converting normal speech to dysarthric speech. N2D VC could be applied in informed decision making related to the medical conditions at the root of the speech pathology. For instance, an oral cancer surgery re-

This chapter has been published as: Huang, W. C., Halpern, B. M., Violeta, L. P., Scharenborg, O., & Toda, T. (2022, May). Towards identity preserving normal to dysarthric voice conversion. In ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP) (pp. 6672-6676). IEEE. The PhD candidate contributed to the writing, the idea, and the evaluation of the project.



(a) Previous works. Left: [5]. Right: [4]



(b) Proposed two-stage approach.

Figure 6.1: Illustration of previous work and the proposed method for N2D VC.

sults in changes to a speaker's voice. The availability of a VC model that can generate how the voice could sound after surgery could help the patients and clinicians make informed decisions about the surgery and alleviate the stress of the patients. Another application is the improvement of automatic speech recognition (ASR) by augmenting the training dataset with additional pathological data. Such augmentation could ease the low-resource constraints of a pathological ASR task.

In addition to the requirements for conventional VC, N2D VC has its own unique requirements, each corresponding to one research question:

RQ1: Do the converted samples sound as natural as real dysarthric samples? Naturalness is a basic requirement in all speech synthesis tasks, but it becomes challenging under the context of N2D VC because listeners seem to confuse *naturalness* and *severity* [4].

RQ2: Is the VC model able to retain the speaker identity of the source normal speaker? Since it is often impossible to collect ground truth pathological speech data of a normal source speaker, training a VC model that directly maps a normal source speech to its pathological counterpart is unfeasible. Thus, specific techniques need to be developed to tackle this issue. In addition, evaluation of similarity is hard because listeners have to determine the similarity of a converted pathological speech to the source speaker while having access to only a normal speech of him/her.

RQ3: Is the VC model able to model severity characteristics in a linear way, so that expert listeners perceive more severe samples as more severe? As the condition of patients deteriorates, the severity of the patient's voice will increase. To capture the progress, it is essential to correctly model the severity of the converted speech. This requires modifying specific attributes of speech, such as speaking rate and insertion of pauses.

In this work, we aim to create an identity preserving N2D VC system. The key advan-

tage of this approach is that it allows arbitrary inputs from the source normal speaker, while preserving its identity. The aim of this work is to evaluate the model in a more practical setting than [4] by taking normal speech as input, which alleviates the need of maintaining a pathological voice bank described there. Inspired by [6], the proposed method is a two-stage approach, as depicted in Figure 6.1b. In the first stage, to capture the unique temporal structure of dysarthric speech, we adopt the Voice Transformer Network (VTN) [1, 7], a sequence-to-sequence (seq2seq) VC model based on the Transformer [8] architecture. The converted speech at this stage has the characteristics of dysarthric speech, with an unwanted speaker identity of the reference dysarthric speaker. Then, the normal source speaker identity is restored through a frame-by-frame autoencoder-based VC model [9], which is assumed to be able to preserve local speech attributes related to dysarthria. We evaluated the proposed method on UASpeech [10], and the method achieves good naturalness results, is able to mimic the severity of pathological speech according to three speech language pathologists, while having limited ability to preserve the source speaker's characteristics.

6.2. RELATED WORKS

6.2.1. NORMAL-TO-DYSARTHRIC VC FOR DATA AUGMENTATION IN ASR

Previous research on data augmentation for dysarthric speech has shown promising improvements in ASR word error rates. The mainstream is to use frame-wise models such as deep convolutional general adversarial networks (DCGANs) [11] or Transformer Encoders [12] to convert the speech timbre. As these models do not change the length, extra procedures are needed to change the speaking rate, including speed perturbation [11] or dynamic time warping [12].

There are several downsides to this line of work. As ASR only requires the various dysarthric features to be modelled, the speaker identity of the normal speaker is not retained after conversion. Also, no evaluation methods were conducted to measure the severity of the samples, which means that it was not verified whether the proposed methods were truly able to model the dysarthric features well. In this work, we use a seq2seq model to jointly convert the timbre and speaking rate, which was shown to be more effective than converting them separately in conventional VC [13]. We also address the identity preservation issue with the proposed two-stage approach and conduct subjective evaluations to verify if the severity is indeed modelled.

6.2.2. NORMAL-TO-DYSARTHRIC VC FOR CLINICAL USAGE

There are two previous works that focus on VC for clinical usage. The diagram on the left of Figure 6.1a depicts an N2D VC system presented in [5], which was a combination of a CycleGAN-based frame-wise VC model and a PSOLA-based speech rate modification process. This method suffers from the same issues as those in Section 6.2.1, including audible vocoder artefacts brought by the extra PSOLA operation, and the inability to preserve the speaker identity of the control speaker.

A different work [4] is depicted on the righthand side of Figure 6.1a. The authors focused on dysarthric-to-dysarthric VC, by using a frame-wise VC model called HL-VQ-VAE [14]. However, the setup was not flexible in that (1) a severity-matched VC setup was

required to avoid the need of varying speech rates, and (2) the method required a pathological source utterance, where in real-world applications we might want to synthesise an arbitrary utterance from the normal source speaker.

6.3. PROPOSED FRAMEWORK

Given a speech sample from a normal speaker, N2D VC aims to change the characteristics into that of a dysarthric speech, while preserving the speaker identity of the source normal speaker. In the following subsections, we describe the two components, the parallel seq2seq model and the nonparallel frame-wise model, of our proposed two-stage approach for N2D VC in detail.

6.3.1. MANY-TO-ONE SEQ2SEQ MODELLING

The goal in the first stage is to completely capture the characteristics of the dysarthric speech. Following [6], we adopted the VTN [1, 7], a Transformer-based [8] seq2seq model tailored for VC. When a parallel corpus is available, seq2seq modelling is considered state-of-the-art due to its ability to convert the prosodic structures in speech, which is critical in N2D VC. However, collecting a parallel corpus is especially difficult in our case since it is impractical (almost not feasible) to collect a large amount of data from dysarthric patients. To solve the data scarcity problem, we applied two techniques, as described below.

First, a TTS pretraining technique is applied which facilitates the core ability of a seq2seq VC model, i.e., encode linguistic-rich hidden representations by pretraining using a large-scale TTS dataset [1, 7]. This technique is flexible in that the VC corpus and the pretraining TTS dataset can be completely different in terms of speaker and content, even when trained between normal and dysarthric speakers. In [6], it was shown that training using only 15 minutes of speech from each speaker can yield good results.

Second, we trained the VTN in a many-to-one (referred to as M2O) fashion. Considering that it is easier to collect data from normal speakers rather than patients, we assume that apart from the data of the source normal speaker, we also have access to a set of parallel training set from multiple normal speakers. Given a training utterance from any of the normal speakers, the VTN model is trained to convert to the predefined target dysarthric speaker. M2O training was also used in [15], except they used an auxiliary phoneme recognition regularisation loss.

6.3.2. NONPARALLEL FRAME-WISE MODEL

In the second stage, given the converted dysarthric speech, the goal is to restore the identity of the source normal speaker while preserving the dysarthric attributes. We adopted the same assumption as in [6]: a nonparallel frame-wise VC model changes only time-invariant characteristics such as the speaker identity, while preserving time-variant characteristics, such as the pronunciation. As in [6], we used *crank* [9], an open-source VC software that combines recent advances in VQVAE [16]-based VC methods, including the use of hierarchical architectures, cyclic loss and adversarial training, to carry out the conversion of the speaker identity step. For the remainder of this paper, we refer to this model as *VAE* for short.

6.4. EXPERIMENTAL SETUP

6.4.1. DATASET

We used the UASpeech dataset [10], which contains parallel word recordings of 15 dysarthric speakers and 13 normal control speakers. The training and test set consist of 510 and 255 utterances, respectively. Each dysarthric speaker is categorized to one of three intelligibility groups: low, mid, and high, which correspond to $0-25\%$, $25-75\%$, and $75-100\%$ subjective human transcription error rate (STER). The intelligibility of each speaker was judged by 5 non-expert American English native speakers. We chose two dysarthric speakers from each intelligibility group (high: M08, M10; mid: M05, M11; low: M04, M12) as test speakers for VC. For each dysarthric speaker, a separate VTN was trained using the data of that speaker and all control speakers. For the VAE model, in our preliminary experiments, we found that it was crucial to train with only the normal data rather than training with a mix of dysarthric and normal datasets. We thus used data from the 13 control speakers only.

6.4.2. IMPLEMENTATION

The implementation of the VTN (the left rounded rectangle in Figure 6.1b) was based on the open-source toolkit ESPnet [17, 18]. The detailed configuration can be found online[1]. The TTS pretraining was conducted with M-AILABS judy [19], which was 31 hr long. *crank*, which we base our implementation of VAE on (the right rounded rectangle in Figure 6.1b), is also open-sourced and can be accessed freely[2]. Parallel WaveGAN (PWG) [20] was used as the neural vocoder. We followed an open-source implementation[3]. The training data of PWG contained the audio recordings of all control speakers in UASpeech.

6.4.3. OBJECTIVE EVALUATION METRICS

The speech sample outputs of the two stages (VTN, VTN-VAE) are separately evaluated using the metrics described in this section, whenever the evaluation does not require ground truth. In this evaluation, we considered conversion pairs between all 13 normal source speakers and the 6 dysarthric speakers mentioned in Section 6.4.1.

P-ESTOI/P-STOI

P-ESTOI/P-STOI were previously demonstrated to work well for the objective evaluation of dysarthric speech [21]. These methods focus on quantifying distortion in the time-frequency structure of the speech signal, which is related to severity and naturalness (RQ1 and RQ3). In short, we used multiple gender-specific ground truth control utterances to form a reference utterance. By calculating the frame-level cross-correlation of each pathological utterance with the reference utterance, we obtain an utterance-level P-ESTOI/P-STOI score. Taking the mean of each utterance-level score, we obtain a speaker-level score, which is correlated with the STER scores for the six speakers to obtain r. This is repeated with the ground truth speakers to obtain r_{GT}

[1] https://github.com/espnet/espnet/tree/master/egs/arctic/vc1
[2] https://github.com/k2kobayashi/crank
[3] https://github.com/kan-bayashi/ParallelWaveGAN

PHONEME ERROR RATE

The phoneme error rate (PER) calculated with a phoneme recognizer evaluates the intelligibility, which is also related to severity and naturalness (RQ1 and RQ3). We use a pre-trained Kaldi ASR model with the same specifications as the one used in [22] for phoneme recognition. The ASR was trained with the TIMIT dataset and used an HMM acoustic model. The TIMIT corpus is an English read speech corpus specifically designed for acoustic-phonetic studies [23]. To measure the PER, we require phonemic transcriptions of the UASpeech utterances (reference). We used g2p-en[4] for grapheme-to-phoneme conversion. The reference is compared to the VC utterances transcribed by the trained ASR.

6.4.4. SUBJECTIVE EVALUATION PROTOCOLS

Subjective evaluation was carried out by naive listeners to assess the naturalness and similarity of samples (RQ1, RQ2). An additional evaluation was done by expert listeners to assess severity (RQ3). Contrary to the objective evaluations, we did not consider all conversion pairs (due to constraints in time and budget). Audio samples can be found online[5].

SEVERITY

We designed an AB evaluation study for evaluating severity (RQ3). In the study, 3 trained speech-language pathologists (SLPs) were asked to listen to two different synthesised utterances from two unknown speakers who have different speech severity and select the synthesised speech sample that they perceived as being more pathological. We used four speaker pairs (see Table 6.4), two for each severity level. For each speaker pair, 20 utterances were rated. After rating the synthesised pathological speech samples, the experiment was repeated with the ground truth samples – as a control for cases where we observe a reversal in the expected severity judgement in the VC speech samples. So, in total, each SLP was asked to rate 80 utterances. A binomial test is performed to calculate significance.

NATURALNESS

In order to evaluate naturalness (RQ1), we followed the setup in [4] with a few modifications based on our previous findings. In our previous study, listeners rated the severity of the speech samples (instead of the naturalness) on a 5-point mean opinion score (MOS) scale. The results showed a flooring effect. Therefore, in this experiment, we increase the resolution of the MOS-scale to have increments of 0.5. The questionnaire starts with an explanation of what we mean with naturalness, followed by an example of natural, normal and pathological (low severity) speech. The respondents were instructed to rate these both as 5 (highly natural). The stimuli consisted of 13 utterances for both pathological speakers of each severity (low, high, mid), leading to a total of 78 utterances. Subsequently, the experiment was repeated with the ground truth samples. The utterances were rated by 30 native American English listeners. A Wilcoxon signed-rank test is performed to calculate significance.

[4] https://github.com/Kyubyong/g2p
[5] https://unilight.github.io/Publication-Demos/publications/n2d-vc

Table 6.1: Objective evaluation results.

	High		Mid		Low			
	M08	M10	M05	M11	M04	M12	r	r_{GT}
P-STOI VTN	0.73	0.75	0.62	0.60	0.58	0.45	0.88	0.89
P-ESTOI VTN	0.37	0.37	0.20	0.16	0.09	0.08	0.93	0.90
P-STOI VTN-VAE	0.73	0.75	0.62	0.63	0.61	0.45	0.84	0.89
P-ESTOI VTN-VAE	0.37	0.35	0.21	0.19	0.12	0.06	0.94	0.90
PER VTN	58.7	55.1	84.1	71.8	79.6	103.4	0.83	0.70
PER VTN-VAE	62.9	59.3	106.3	76.2	81.2	120.0	0.68	0.70
STER	7.0	7.0	42.0	38.0	98.0	92.6	1.0	–

Table 6.2: Mean opinion score results of the naturalness test with 95% confidence intervals. Columns correspond to the intelligibility level, and rows correspond to ground truth (GT) and synthetic (VC) results. Higher is better.

	Normal	High	Mid	Low
GT	3.93 ± .54	3.92 ± .54	2.86 ± .89	2.32 ± 1.16
VC	-	2.70 ± .95	2.28 ± 1.03	1.94 ± 1.21

SIMILARITY OF THE VOICE WITH THE SOURCE NORMAL SPEAKER

For the similarity (RQ2) evaluation, we follow the protocol in [4]. Listeners are presented a converted sample and a reference sample, and are asked to judge whether the two samples are uttered by the same speaker. In short, the evaluation is AB similarity study where the source speaker is a pathological speaker, the target speaker is the control speaker. The reference speech is either from the source (Similarity to source) or the target (Similarity to target). We selected three pathological speakers (M04, M11, M10) which have deemed to have recognisable characteristics in our previous study [4]. Furthermore, we randomly sampled (without replacement) two control speakers for each pathological speaker. The test were done by 5 naive American English listeners. A binomial test is performed to calculate significance.

6.5. EVALUATION RESULTS

6.5.1. OBJECTIVE EVALUATIONS

P-STOI/P-ESTOI

The second block of Table 6.1 summarises the results of the P-STOI/P-ESTOI analyses. In the VTN stage, the obtained correlation between the P-STOI/P-ESTOI and the STER are similar to the ones one would obtain with the GT (r_{GT}). Therefore, in the VTN stage the severity is well captured. In the VTN-VAE stage, the P-STOI correlation decreases from 0.88 to 0.84, while the P-ESTOI slightly increases from 0.93 to 0.94, which is a bit higher than (r_{GT}). This latter change can be explained as follows: the frame-based VAE model does not change the temporal aspects of the signal but rather the spectral aspects, for

Table 6.3: Results of the similarity AB experiments with 95% confidence intervals.

	Similarity to target	Similarity to source
M04→CM05	20% ± 10%	32% ± 12%
M11→CM09	37% ± 13%	43% ± 13%
M10→CF03	55% ± 13%	8% ± 7%
M04→CM04	33% ± 12%	27% ± 12%
M11→CM10	23% ± 11%	32% ± 12%
M10→CF02	48% ± 13%	10% ± 8%
Ideal	100%	0%

Table 6.4: Percentage of "correct" answers in the AB severity tests for the ground truth samples and the different stages of the architecture. *** is $p < 0.001$; * $p < 0.05$

Speaker pairs	**Ground truth**	**VTN**	**VTN-VAE**	**Severity**	
M04 vs M05	95% ***	85% ***	53%	Low	Mid
M05 vs M08	90% ***	95% ***	80% ***	Mid	High
M12 vs M11	93% ***	85% ***	75% *	Low	Mid
M11 vs M10	98% ***	95% ***	68% *	Mid	High

which the P-ESTOI has a higher sensitivity.

PHONEME ERROR RATE

The VTN PER results in Table 6.1 show higher correlation with the STER than the GT, which indicates that we can mimic the severity aspects of the pathological speech in the first stage. However, PER VTN-VAE results are decreased compared to the PER VTN. This is probably because the VAE stage causes a naturalness degradation.

6.5.2. SUBJECTIVE EVALUATIONS

NATURALNESS

Table 6.2 shows the MOS results. First, similar to our previous study [4], we observe that with decreasing intelligibility, naive listeners perceive the heard speech increasingly unnatural – even in the case of ground truth samples. Second, the ground truth samples are consistently rated as more natural than the converted ones ($p < 0.001$). Although, these results are not directly comparable to [4], we note that we've observed overall higher MOS values. We suggest that the use of seq2seq models contributed to this improvement, and such quality is sufficient for further investigation.

SIMILARITY

Table 6.3 describes the identity preservation ability of the VC framework. We can see that the Similarity to source column has less than 50% similarity for all speaker pairs, therefore we can conclude that the VC can successfully ignore the pathological source speaker's characteristics. However, we can also see from the Similarity to Target column that (except from the M10→CF03) none of the VC samples have more than 50% similarity to the target. Such results emphasise the "unobtainable ground truth" difficulty faced by

the model, as described in Section 6.1. Meanwhile, this also points out that improving speaker similarity is an important future work, as this problem was also present in [6].

SEVERITY

Table 6.4 lists the percentage of "correct" answers in the AB severity test done with the SLPs. On average, the SLPs always perceived the more severe speakers as more severe (each entry in Table 6.4 is over 50%). In the first VTN stage, no more than 10% decrease in "correct" answers is observed in the severity recognition compared to the ground truth. Furthermore, the ratio of "correct" severity decisions slightly increased in the case of the VTN M05 vs M08 pair. This indicates that the VTN simulates the severity aspects well. After the second VTN-VAE step, we see a decrease in "correct" answers, which means that in the case of speaker-specific samples, the SLPs made more errors in indicating which of the two samples had a worse severity, possibly because the severity difference is less perceivable for the SLPs due to the additional distortion caused by the VAE.

6.6. CONCLUSIONS

In this paper, we proposed a novel two-stage framework for N2D VC. We evaluated the proposed method on UASpeech [10], and the method achieved good naturalness results, was able to mimic the severity characteristics in a linear way according to three speech language pathologists, while being able to convert away from the pathological source speaker's characteristic. In the future, we will focus on improving the preservation of the normal source speaker identity.

6.7. ACKNOWLEDGEMENTS

We would like to thank Lisette van der Molen, Klaske van Sluis, and Marise Neijman for participating in the severity experiment. All questionnaire participants were remunerated justly (7.50GBP/hour) in each experiment. B.M.H. is funded through the EU's H2020 research and innovation programme under MSC grant agreement No 766287. The Department of Head and Neck Oncology and Surgery of the Netherlands Cancer Institute receives a research grant from Atos Medical (Hörby, Sweden), which contributes to the existing infrastructure for quality of life research. This work was partly supported by JSPS KAKENHI Grant Number 21J20920, JST CREST Grant Number JPMJCR19A3, and AMED under Grant Number JP21dk0310114, Japan.

BIBLIOGRAPHY

[1] W.-C. Huang, T. Hayashi, Y.-C. Wu, H. Kameoka, and T. Toda, "Voice transformer network: Sequence-to-sequence voice conversion using transformer with text-to-speech pretraining," in *Proc. Interspeech*, 2020, pp. 4676–4680.

[2] K. Zhou, B. Sisman, and H. Li, "Transforming Spectrum and Prosody for Emotional Voice Conversion with Non-Parallel Training Data," in *Proc. Odyssey*, 2020, pp. 230–237.

[3] M. Cotescu, T. Drugman, G. Huybrechts, J. Lorenzo-Trueba, and A. Moinet, "Voice conversion for whispered speech synthesis," *IEEE Signal Processing Letters*, vol. 27, pp. 186–190, 2019.

[4] M. Illa, B. M. Halpern, R. van Son, L. Moro-Velazquez, and O. Scharenborg, "Pathological voice adaptation with autoencoder-based voice conversion," in *Proc. SSW11*, 2021, pp. 19–24.

[5] B. M. Halpern, J. Fritsch, E. Hermann, R. van Son, O. Scharenborg, and M. Magimai-Doss, "An objective evaluation framework for pathological speech synthesis," in *Speech Communication; 14th ITG Conference*. VDE, pp. 1–5.

[6] W.-C. Huang, K. Kobayashi, Y.-H. Peng, C.-F. Liu, Y. Tsao, H.-M. Wang, and T. Toda, "A Preliminary Study of a Two-Stage Paradigm for Preserving Speaker Identity in Dysarthric Voice Conversion," in *Proc. Interspeech*, 2021, pp. 1329–1333.

[7] W. C. Huang, T. Hayashi, Y. C. Wu, H. Kameoka, and T. Toda, "Pretraining techniques for sequence-to-sequence voice conversion," *IEEE/ACM TASLP*, vol. 29, pp. 745–755, 2021.

[8] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin, "Attention is All you Need," in *Proc. NIPS*, 2017, pp. 5998–6008.

[9] K. Kobayashi, W.-C. Huang, Y.-C. Wu, P. L. Tobing, T. Hayashi, and T. Toda, "crank: An open-source software for nonparallel voice conversion based on vector-quantized variational autoencoder," in *Proc. ICASSP*, 2021, pp. 5934–5938.

[10] H. Kim, M. Hasegawa-Johnson, A. Perlman, J. Gunderson, T. S. Huang, K. Watkin, and S. Frame, "Dysarthric speech database for universal access research," in *Proc. Interspeech*, 2008, pp. 1741–1744.

[11] Z. Jin, M. Geng, X. Xie, J. Yu, S. Liu, X. Liu, and H. Meng, "Adversarial Data Augmentation for Disordered Speech Recognition," in *Proc. Interspeech 2021*, 2021, pp. 4803–4807.

[12] J. Harvill, D. Issa, M. Hasegawa-Johnson, and C. Yoo, "Synthesis of New Words for Improved Dysarthric Speech Recognition on an Expanded Vocabulary," in *Proc. ICASSP*, 2021, pp. 6428–6432.

[13] J. Zhang, Z. Ling, L. Liu, Y. Jiang, and L. Dai, "Sequence-to-Sequence Acoustic Modeling for Voice Conversion," *IEEE/ACM TASLP*, vol. 27, no. 3, pp. 631–644, 2019.

[14] T. V. H. and M. A., "Non-parallel Voice Conversion based on Hierarchical Latent Embedding Vector Quantized Variational Autoencoder," in *Proc. Joint Workshop for the Blizzard Challenge and Voice Conversion Challenge 2020*, 2020, pp. 140–144.

[15] F. Biadsy, R. J. Weiss, P. J. Moreno, D. Kanvesky, and Y. Jia, "Parrotron: An End-to-End Speech-to-Speech Conversion Model and its Applications to Hearing-Impaired Speech and Speech Separation," in *Proc. Interspeech*, 2019, pp. 4115–4119.

[16] A. v. d. Oord, O. Vinyals, and K. Kavukcuoglu, "Neural Discrete Representation Learning," in *Proc. NIPS*, 2017, pp. 6309–6318.

[17] S. Watanabe, T. Hori, S. Karita, T. Hayashi, J. Nishitoba, Y. Unno, N. E. Y. Soplin, J. Heymann, M. Wiesner, N. Chen, A. Renduchintala, and T. Ochiai, "ESPnet: End-to-End Speech Processing Toolkit," in *Proc. Interspeech*, 2018, pp. 2207–2211.

[18] S. Watanabe, F. Boyer, X. Chang, P. Guo, T. Hayashi, Y. Higuchi, T. Hori, W.-C. Huang, H. Inaguma, N. Kamo, S. Karita, C. Li, J. Shi, A. S. Subramanian, and W. Zhang, "The 2020 ESPnet Update: New Features, Broadened Applications, Performance Improvements, and Future Plans," in *Proc. IEEE Data Science and Learning Workshop (DSLW)*, 2021, pp. 1–6.

[19] Munich Artificial Intelligence Laboratories GmbH, "The M-AILABS speech dataset," 2019, accessed 30 November 2019. [Online]. Available: https://www.caito.de/2019/01/the-m-ailabs-speech-dataset/

[20] R. Yamamoto, E. Song, and J. Kim, "Parallel WaveGAN: A fast waveform generation model based on generative adversarial networks with multi-resolution spectrogram," in *Proc. ICASSP*, 2020, pp. 6199–6203.

[21] P. Janbakhshi, I. Kodrasi, and H. Bourlard, "Pathological speech intelligibility assessment based on the short-time objective intelligibility measure," in *Proc. ICASSP*, 2019, pp. 6405–6409.

[22] M. Purohit, M. Patel, H. Malaviya, A. Patil, M. Parmar, N. Shah, S. Doshi, and H. A. Patil, "Intelligibility improvement of dysarthric speech using mmse discogan," in *International Conference on Signal Processing and Communications (SPCOM)*, 2020, pp. 1–5.

[23] J. S. Garofolo, L. F. Lamel, W. M. Fisher, J. G. Fiscus, and D. S. Pallett, "DARPA TIMIT acoustic-phonetic continous speech corpus CD-ROM. NIST speech disc 1-1.1," *NASA STI/Recon technical report n*, vol. 93, p. 27403, 1993.

7

DISCUSSION AND CONCLUDING REMARKS

In this thesis, we have presented a series of studies on three use cases of accessible speech technology: automatic speech recognition, automatic speech severity evaluation and voice conversion. In this final chapter, we are going to revisit the three main research questions in light of the experimental findings presented in the main body of the thesis.

7.1. RQ1: ON THE SOURCES OF BIAS IN ATYPICAL AND PATHOLOGICAL SPEECH

Automatic speech recognition currently does not work equally well for all users of speech technology. Many studies have shown that ASR performance is diminished for users with an accent [1], dialect or speech pathology [2]. When we are talking about performance degradation of a speaker group (e.g. pathological speakers) compared to the typical user of a speech technology application, we say that there is a bias against that first speaker group. Our first research question is related to the extent and different sources of bias. We quantified the bias against speakers of Dutch, Mandarin Chinese (Chapter 2); and American English speakers with oral cancer (Chapter 3). Furthermore, we investigated possible sources of this bias, specifically, how pronunciation, the severity of speech, and external noise affects the bias.

The studies in Chapters 2 and 3 have demonstrated that automatic speech recognisers are biased against atypical and pathological speakers. First, in Chapter 2, we found significant bias against (Dutch) male speech, children's speech, old adults' speech, non-native speech, and (Dutch and Mandarin) regionally accented speech. Furthermore, in Chapter 3, we found that there is a bias against oral cancer speech, showing that the more severe the oral cancer speech is, the higher the word error rates are.

Our studies first looked into pronunciation as a possible source of bias. To analyse the impact of pronunciation on recognition errors, we analysed phoneme error rates

(PERs). Suppose we observe that the ASR has low recognition rates on a particular phoneme. This might indicate that the particular phoneme deviates significantly from the typical phoneme production, e.g. due to impaired articulation or low language proficiency.

In Chapter 2, we calculated the PERs of several atypical speaker groups in Dutch and Mandarin and compared them to each other. The comparison showed that the phonemes with the highest PERs were more or less consistent for most speaker groups, e.g. in Dutch /ʃ/ was among the top-5 highest PER phonemes for nearly all speaker groups; in Mandarin, /ʐ̩/ was always the most problematic. Furthermore, we found occasionally that phonemes that are known to pose difficulties for the speaker groups were also recognised with a high PER. Examples include /œy/ and /y/ in the case of Dutch non-native speakers [3] or the variation of /s/ and /ʂ/ in the case of Min and Xiang regional dialects of Mandarin Chinese. However, in general, we found that phonemes with the highest PERs were the phonemes with the lowest frequency in the training material, e.g., the /ʃ/, /ɲ/, /ʒ/ for Dutch.

In Chapter 3, we looked into whether pathological pronunciations affect bias rates through a similar phoneme error analysis as in Chapter 2 and an additional articulatory feature error analysis. The study found that plosives and some vowels (/aa/ and /uw/) were challenging to recognise for the ASR system, with /g/ and /p/ having a PER of over 60%. As these are sounds that are known to be degraded in the case of oral cancer [4, 5], we can state with high confidence that oral cancer pronunciation affects bias. Therefore, we conclude that pronunciation differences between typical and atypical speakers are a source of bias.

Another possible source of bias could be the different noise levels for atypical and typical speakers. Speech recognisers are often used on smartphones, which can vary significantly in terms of built-in microphone quality, with cheaper smartphones having typically poorer microphone quality [6]. Chapter 3 used a oral cancer speech corpora collected from YouTube, therefore we were interested whether the different noise levels in the corpora affected our results. We investigated this potential source of bias by comparing the signal to noise ratio of the speakers' recordings, the speech severity of the oral cancer speakers, and the WER of the oral cancer speakers' recordings in Chapter 3. The comparison showed that the WER was not correlated with the signal to noise ratio but highly correlated with the perceived severity of the speech. While it is evident that external noise is a factor in general ASR performance [7], in the case of pathological speech, it was clearly the severity of the speech pathology that had the greatest impact on the bias.

During the phoneme analysis in Chapter 2, we observed that children's speech is annotated in a more lenient way than those of non-native adult speakers. For example, we found more restarts with non-native adults (e.g., annotator used `uh probier uh probeert` while simply `probier` with children) than in non-native children. We also saw that for non-native speakers, sometimes words were not annotated with their standard spellings, causing an out of vocabulary issue, therefore increasing the word error rate (e.g., `brie-ven-bus` in non-native speakers while `brievenbus` with children). These differences in annotation are often not justified by the acoustic differences. These clues point to the annotation and the annotation instructions as a possible source of bias [8, 9]. Finally, the choice of ASR architectures also influenced the observed bias. For instance, in Chapter 2, we found a larger bias for the end-to-end models against male speakers for

Dutch teenagers and older adults, non-native accents, against Flemish, and observed more bias against more strongly accented Mandarin.

There are many promising avenues to alleviate the issues mentioned above and reduce bias. To mitigate bias in general, the most important is to balance and diversify the types of speakers in the training and evaluation data of ASR models, as shown in Chapter 2 and [10, 11]. Unfortunately, balancing the dataset regarding the types of speakers is not always straightforward since atypical speech is often scarcer than typical speech. Data augmentation could be a way to increase the amount of training data for the under-resourced speaker group and reduce bias, as shown in our recent work [12].

To mitigate bias due to pronunciation and speech severity, we suggest using fMLLR. In Chapter 3, fMLLR achieved both the best test set performance and the lowest correlation between the speaker WERs and the speaker severity, meaning that it is the least biased by severity. Our finding is consistent with other studies recommending fMLLR for pathological speech recognition [13, 14, 15, 16]. In recent work (not part of this thesis) [12], Domain Adversarial Training (DAT) was shown to be successful in reducing bias against Dutch non-native accented speech. DAT is a training method which aims to extract acoustic features that are invariant for specific characteristics (e.g. accent) of the speech. Potentially, it can also be used to reduce bias against pathological speech.

7.2. RQ2: ON SEVERITY EVALUATION OF SPONTANEOUS SPEECH

Most oral cancer patients have pathological speech side effects after the treatment of oral cancer. To address these side effects, oral cancer patients require speech therapy supervised by a speech-language pathologist. As we want to measure the efficiency of speech therapy, we require a way to track the progress during speech therapy. To track progress during speech therapy, subjective questionnaires are typically used to estimate the severity of the speech at regular intervals [17]. However, these subjective questionnaires are often criticised as they can be highly unreliable [18]. Therefore, a significant amount of research is directed towards the objective estimation of speech severity [19]. However, there are multiple shortcomings of these objective approaches. First, all of these approaches use read speech in their evaluation, while spontaneous speech would more accurately reflect the patients' communicative issues. We call this issue an issue of *ecological validity*. Second, clinicians want to make sure that objective approaches estimate severity based on pathology-specific evidence (e.g. breathiness) and not on extralinguistic cues (e.g. accent). In other words, there would be a need for more estimation models that are *explainable*. To summarise, our second research question concerned the automatic prediction of severity ratings in the case of spontaneous oral cancer speech with explainable models. This research question was studied in Chapter 4.

Our main finding was that objective methods correlate highly with the subjective severity evaluations of SLPs on our spontaneous oral cancer corpora. Specifically, we investigated two sets of approaches. The first set of approaches was called the reference-free approaches. Reference-free approaches estimated the speech severity solely based on the acoustics of the speech signal. The second set of approaches was called the reference-based approaches. Contrary to the reference-free approaches, reference-based approaches also use a reference (e.g. a transcription of the speech signal or a parallel healthy speech signal) in the severity estimation. The expectation was that reference-

based approaches would work better than reference-free approaches as these can use more information. However, transcription of speakers with low intelligibility is sometimes difficult or even impossible to obtain, meaning that reference-free approaches still have merit.

We evaluated these two sets of approaches on two evaluation conditions called reference-based and reference-free evaluation conditions. In the first, reference-based evaluation condition, we evaluated the approaches on transcribed utterances of the speakers. In that evaluation, we found that the best method was an ASR-based technique. Furthermore, our detailed comparison of the ASR experiments showed that ASR models which used oral cancer speech during their training produced higher correlations with the subjective severity scores. We think that adding some oral cancer data to the training material is beneficial to the ASR models because the acoustic models then capture some of the mild disfluencies due to oral cancer speech, similar to how human listeners quickly adapt to mild disfluencies in healthy speech [20, 21].

In the second, reference-free evaluation condition, the approaches were evaluated on all utterances, including utterances without transcriptions. We also set up the evaluation so that all utterances of a recording were used to estimate the severity instead of a single utterance from a recording. This reference-free evaluation found that the best approaches used modulation spectrum and global variance acoustic features. These acoustic features are often used to evaluate synthetic speech samples in VC and text-to-speech (TTS) synthesis [22]. This finding shows that acoustic measures used for the evaluation of the naturalness of synthetic speech can also be used to evaluate speech severity.

The second part of our research question concerned the explainability of automatic speech severity evaluation models. In Chapter 4, we saw that the explainable models attained high correlations with the subjective scores on the reference-free evaluation. It is thus possible to build automatic severity evaluation systems that perform well and are explainable. Chapter 4 focused on the performance of these models and did not go into detail regarding the explanation process itself. However, our previous work [23] (not part of this thesis) showed more of this explanation process. All of the reference-free approaches in Chapter 4 used the LASSO model, which is a variant of linear regression. Linear regression is already considered an explainable machine learning model - the influence of the individual acoustic features on the prediction can be understood by looking at the regression coefficients. [23] showed that LASSO further distils the predictions it makes by finding a prediction model that has fewer non-zero coefficients. In other words, LASSO finds a model that uses only a subset of the acoustic features for the decision. Even if the models are explainable, we also need to ensure that the acoustic features are explainable to make the entire decision process explainable. For example, if the sixth global variance dimension is the most important for the estimation, we can figure out which frequency variations the sixth Mel-frequency cepstral coefficients correspond to. From the frequency variations, we can explain what kind of acoustic-phonetic cues the severity decision is based on.

Automatic severity evaluation techniques could be improved in the future. The findings of the reference-based and reference-free approaches both show possible avenues for improvements. One finding in the reference-based approaches was that adding a

certain amount (here: around 1 hour) of oral cancer training data is beneficial for automatic severity evaluation. However, it is still not clear how much that "certain amount" is in general. We think that this would warrant a further set of experiments where we could determine the optimal ratio of healthy and pathological speech data in ASR training for the goal of speech severity evaluation. Furthermore, our findings regarding the reference-free approaches showed that acoustic features used for evaluating synthetic speech in VC and TTS are also useful for evaluating pathological speech severity. It follows that these acoustic features are currently capturing a common set of acoustic cues related to the naturalness and severity of speech (e.g. modulation in frequency and lack of variability). Most likely, these common acoustic cues correspond to a shared property of these concepts, i.e., both naturalness and severity express a distortion of the speech. Therefore, it follows that a better understanding of what is the severity (and naturalness) apart from distortion would likely lead to better automatic severity evaluation techniques.

7.3. RQ3: ON THE CONVERSION OF HEALTHY VOICES TO PATHOLOGICAL VOICES

Patients undergoing oral cancer surgery want to know what they might sound like after surgery. It would be important to procure a tool that can show an example of speech-related side effects after oral cancer treatment. A possible tool would be based on voice conversion, which we have already described in Section 1.4.3. Voice conversion could convert healthy speech into pathological speech to show examples of speech-related side effects. Our third research question concerned whether it is possible to build a voice conversion system that preserves the identity of the healthy speaker while achieving naturalness that is comparable to real oral cancer speech.

There are two key difficulties with this voice conversion task. The first difficulty is that speech pathology affects both the speaker's voice characteristics (speaker identity) and the speaker's intelligibility. In voice conversion, it is usually either the intelligibility (e.g., cross-lingual voice conversion) or the speaker identity that is changed (e.g., voice impersonation). We named this difficulty as *disentanglement* issue. The second difficulty of our voice conversion task is related to the *evaluation* of naturalness and speaker similarity (see Section 1.4.3). Pathological speech is already perceived as unnatural (see Chapter 5), therefore, even if we could perfectly imitate pathological speech using our model, it would affect the naturalness evaluation result.

7.3.1. VOICEBANKING APPROACH

We have attempted to address these issues via the two voice conversion models introduced in Chapters 5 and 6. In Chapter 5, we investigated a non-parallel autoencoder-based approach to convert a pathological speech utterance to the voice characteristics of another speaker based on the speaker's embedding. By using an autoencoder-based approach, we could automatically learn the *disentanglement* between the pathological speaker's identity and pathological speech severity. As we had no access to parallel healthy and dysarthric utterances from the same speaker for evaluation, we evaluated the approach on the task of converting the speech of a dysarthric speaker to the speech

of another dysarthric speaker with similar intelligibility.

By choosing speaker pairs with similar speaker intelligibility, the experiments could ignore the conversion of the intelligibility (time-variant) aspects. Ignoring the intelligibility aspects was important for multiple reasons. First, during the speaker similarity *evaluation*, even if two identical speakers are presented to the listeners, intelligibility differences could lead to reduced speaker similarity results [24]. Second, because the intelligibility (therefore; most likely the severity too) of the source and target speaker is the same, any decrease in the MOS score of the converted sample is due to a loss of naturalness. Even if the intelligibility decreases during conversion, this change is unwanted, contrary to the situation described in Section 1.4.3, therefore it can be regarded as a loss in naturalness.

The evaluation experiments in Chapter 5 demonstrated reasonable naturalness; however, the difference in naturalness between the ground truth and converted samples was always significant. We also found that the less intelligible the pathological speech is, the lower its perceived naturalness. The evaluation experiments also demonstrated successful conversion of speaker identity for high and low intelligibility speakers. For one of the mid intelligibility speakers, we noticed that listeners could not verify the speaker when presented with multiple ground truth recordings from that speaker. This lack of recognition demonstrates that the ill-posed speaker recognition problem mentioned in Section 1.4.3 can indeed be a problem in practice.

The conversion experiments carried out in Chapter 5 were limited in many aspects. First, we used only male speakers in the conversion due to other constraints in our experiment design. Furthermore, a significant disadvantage of the used voice conversion setup is that it needs pathological source speech. This disadvantage is limiting as we cannot synthesise an arbitrary pathological utterance with a new patient's voice. In other words, only utterances that are present in a previously collected pathological voice dataset can be customised to the new patient's voice.

7.3.2. TWO-STAGE APPROACH

Therefore, we wanted to build a voice conversion that can address this limitation and is able to synthesise arbitrary utterances. In Chapter 6, we developed a sequence-to-sequence model for the conversion of healthy speech to pathological speech. During the experimentation, we noticed that this model converted the intelligibility (time-variant) aspects of the pathology well but did not retain the speaker characteristics of the healthy source speaker.

However, having addressed the speaker identity conversion in Chapter 5, we could combine the systems in Chapters 5 and 6 to create a framework which retained the speaker's identity while changing the severity of the speech. The two-pronged approach is as follows: the first sequence-to-sequence system converts healthy speech to speech of a model pathological speaker (changing the time-variant aspects), and then the voice characteristics of this model pathological speaker are customised to the healthy speaker's voice characteristics using the auto-encoder (changing the time-invariant aspects). In contrast to the conventional voice conversion approach, we have now three speakers: the source speaker, the model pathological speaker, and the target speaker. The model pathological speaker can be thought of as a generic pathological speaker who repre-

sents the intelligibility (time-variant) aspects of the pathology well but does not have the speaker characteristics of the healthy speaker.

The two-pronged approach allowed us to check after each stage that the desired properties (the intelligibility and the speaker identity) are converted after the respective stage while being able to synthesise any arbitrary utterance. In the first stage, when we converted to the model pathological speaker, we investigated if the intelligibility aspects were captured well. We ran multiple evaluations to check this aspect: (1) a phoneme recognition task; (2) an objective evaluation with P-STOI and P-ESTOI, which have been demonstrated to estimate the severity of dysarthric speech reliably on multiple dysarthric corpora [25]; and (3) a subjective evaluation with three trained speech language pathologists. These evaluations all showed that the intelligibility/severity aspects of the speech were captured well.

After the second stage of the conversion, we investigated whether we could recover the original, healthy speaker's identity in the conversion. The test results showed that the proposed model was not able to: the converted samples were neither recognised as the healthy speaker nor as to the model pathological speakers. There are multiple possible explanations for this.

The first explanation is that the model cannot convert the speaker characteristics well. This explanation is somewhat surprising, knowing that the model presented in Chapter 5 and the second stage of Chapter 6 are nearly identical. It is possible that the conversion after the first stage has reduced naturalness, which affects the speaker identity conversion in the second stage. It is also possible that other differences in the model architecture led to this difference in performance.

The second reason could be due to the setup of the speaker identity evaluation experiment. In standard speaker identity evaluation, the ground truth samples have the same intelligibility as the converted samples. However, we did not have parallel dysarthric and healthy data from the same speaker, therefore we could not perform the experiments in the standard way. Specifically, the problem was that listeners had to imagine how the control speaker's dysarthric speech would sound when making their judgements about the speaker's identity. Figure 7.1 illustrates that even if we did the evaluation with real samples or the output of a perfect VC system, this evaluation would be challenging for the listeners. We hypothesise that this phenomenon could have affected the speaker identity evaluation results negatively. The speaker identity evaluation results indeed showed a trend that is consistent with this hypothesis - if we used a more intelligible model speaker in the conversion, the converted speakers were deemed more similar to the target speaker and less similar to the source speaker.

We were also interested in how natural the converted samples are compared to the real samples. Similarly to the findings of Chapter 5, we found that less intelligible samples were rated by the listener as less natural. The converted samples had reasonable but still significantly lower naturalness than the pathological ground truth. Compared to Chapter 5, the obtained naturalness results were higher; however, the two experiments had different stimuli composition; therefore, the two evaluations are not directly comparable. We conclude that the voice conversion framework is (1) able to convert the severity/intelligibility aspects of the speech; (2) retains reasonable naturalness after conversion, however, it needs more improvement; (3) does not retain the identity of the

speaker.

7.3.3. FUTURE WORK

To improve the naturalness and speaker identity aspect, the voice conversion based model needs to be improved. Average modelling based approaches seem to be a good candidate, as these have higher naturalness than autoencoder-based approaches in the VCC2020 challenge [26]. A significant challenge in adopting these models for pathological voice conversion is that these models use phonetic posteriorgram (PPG) features. PPG features are extracted from ASRs, which are currently not working well for atypical speech. However, as ASRs get better on atypical speech, they might open a window of opportunity to use these approaches for pathological voice conversion.

Furthermore, there are still some challenges remaining with the *evaluation* of naturalness and speaker identity in voice conversion. First, the experiments in Chapters 5 and 6 showed that speech severity affects the listener's evaluations of the naturalness of the generated pathological speech, which is not ideal because we want to produce pathological speech that is natural and has the right speech severity level. Therefore, instead of relying on subjective measures of naturalness, objective evaluation measurements could be a good alternative. However, a lot of existing objective measures for naturalness are also affected by speech severity (see Chapter 4). Therefore, it would be imperative to either develop objective metrics and subjective evaluation protocols that are only sensitive to the naturalness of the speech, and not the severity of the speech.

Second, as mentioned above, the speaker identity experiments in Chapter 6 seemed to indicate that it is very difficult for naive listeners to recognise the same speaker with different intelligibility, even in ground truth pathological speech. To alleviate the recognition issues, it could be interesting to recruit a patient's family members to evaluate the speaker identity of the generated pathological speech, as it is known that recognition of familiar speakers is much easier and uses different neural pathways than naive speaker recognition [27]. In scenarios where it is possible to collect parallel data, parallel pathological and healthy data could be used in the speaker identity evaluation experiments to establish a baseline of recognisability for the voice conversion task. Establishing such a baseline would not solve the evaluation issue, but it would still allow us to separate the issue of the speaker identity conversion from the issue of the listener recognition.

7.4. CONCLUDING REMARKS

Speech technology progressed remarkably in the recent decade, helping society with everyday tasks. As speech technology gets better, more and more people will consider it as a solution in their products and services. Many speech-driven products are already on the market, however, these do not work for everyone (see Chapter 2). In this thesis, we looked at three speech technology applications and how they worked for atypical and pathological speakers, namely: atypical automatic speech recognition, automatic speech severity evaluation, and pathological voice conversion. In this final section, we would like to summarise the findings and provide our opinion on the issues presented in this thesis.

The findings of Chapter 2 confirm that there is a *bias* against atypical and patholog-

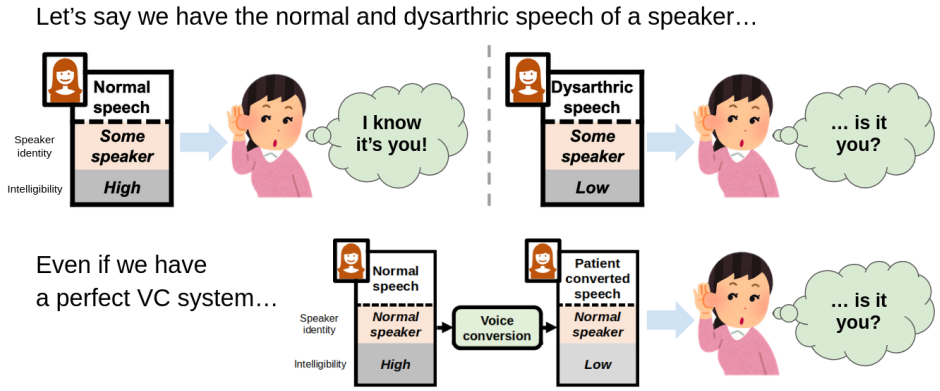


Figure 7.1: Demonstration of why speaker identity evaluation might be difficult in the case of pathological speech.

ical speakers in state-of-the-art automatic speech recognition. Sources of this bias include the composition of the training data, pronunciation and severity of the pathological speech. To mitigate that bias, we should divert our attention from blindly lowering absolute error rates on general test sets as is standard practice, but rather evaluate automatic speech recognition systems on different speaker groups and develop ASR systems that generalise across demographics, dialects, accents and speech pathology.

Regarding the evaluation of speech severity, we have seen that automatic speech recognisers can estimate the severity of oral cancer speech in *ecologically valid* conditions well. A selection of these methods was explainable, and we briefly elaborated on the process of explanation with these models.

Regarding pathological voice conversion, it is possible to convert healthy speech to pathological speech with reasonable naturalness, but speaker-related aspects have to be improved. We presented two voice conversion models for synthesising pathological speech and alleviating the *disentanglement* issue. (1) A VC model that starts from a pathological utterance and uses a speaker embedding to convert it to the style of a different speaker. (2) A two-step approach which converts the intelligibility aspects of the speech first, and then the speaker's identity. Furthermore, we presented two strategies to *evaluate* the naturalness of pathological speech. (1) When converting pathological speech to the pathological speech of the speaker with the same intelligibility, the mean opinion score (MOS) can be used reliably to evaluate naturalness, as no intelligibility degradation is expected. (2) When the severity of the pathological speech is also converted, objective measures should be used alongside subjective raters, as it is clear from our findings that the subjective raters are influenced by the speech severity when rating naturalness.

While the issues (*bias*, *ecological validity*, *disentanglement*, *evaluation*) presented above seem to be distinct, there is a great deal of commonality between them. First, it is exactly automatic speech recognisers' sensitivity to severity of the pathological speech - the *bias* - that makes ASRs a useful tool for the objective evaluation of speech severity. To

put it in different terms, if we did not have high word error rates on pathological speech, ASRs could not be used for severity evaluation. In the severity evaluation scenario, it is desired that ASRs make the right mistakes - mistakes on articulation and not mistakes on accents.

Second, the difficulties of the objective *evaluation* of synthetic pathological speech are similar to the difficulties of the objective evaluation of natural pathological speech. Voice conversion systems are optimised for metrics such as the mean opinion score, modulation spectrum and global variance. If these metrics are sensitive to the severity of the pathological speech, optimisation towards these metrics will lead to enhanced speech rather than pathological speech. The reverse argument also holds for the evaluation of speech severity. If the speech severity metrics are also sensitive to noise artefacts present in synthetic speech, it means that speech severity can only be evaluated in high quality, controlled recording conditions. The requirement of these conditions renders speech severity metrics difficult to use in uncontrolled, spontaneous, *ecologically valid* scenarios.

All of the issues presented above show that the field of speech technology suffers from a blind optimisation to metrics which often do not represent well our desired objectives. In the case of automatic speech recognition, it is the usage of word error rate that needs more careful consideration by evaluating on atypical groups. In the case of voice conversion, it is the mean opinion score which needs to be rethought. In the case of severity evaluation, it is the similarity of naturalness and severity that is problematic. We firmly believe that the solution to these issues lies in understanding the edge cases where these metrics do not work. Understanding these edge cases will lead to new, golden metrics which represent our desired objectives better. Until we have the golden metrics for these tasks, the best we can do is to involve the actual future users of speech technology in the prototyping process as soon as possible. Involving these users is sometimes less than easy (e.g. due to GDPR or medical ethics regulations), but talking with users similar to the target user or having anonymous interactions with the user could be a great substitute. If we can reach out to these potential early adopters of new speech technology applications, we will not only save time but also work on the right research questions, building better and more accessible technologies.

BIBLIOGRAPHY

[1] A. Hinsvark, N. Delworth, M. Del Rio, Q. McNamara, J. Dong, R. Westerman, M. Huang, J. Palakapilly, J. Drexler, I. Pirkin *et al.*, "Accented speech recognition: A survey," *arXiv preprint arXiv:2104.10747*, 2021.

[2] K. Rosen and S. Yampolsky, "Automatic speech recognition and a review of its functioning with dysarthric speech," *Augmentative and Alternative Communication*, vol. 16, no. 1, pp. 48–60, 2000.

[3] A. Neri, C. Cucchiarini, and H. Strik, "Selecting segmental errors in non-native dutch for optimal pronunciation training," 2006.

[4] J. Takatsu, N. Hanai, H. Suzuki, M. Yoshida, Y. Tanaka, S. Tanaka, Y. Hasegawa, and M. Yamamoto, "Phonologic and acoustic analysis of speech following glossectomy

and the effect of rehabilitation on speech outcomes," *Journal of Oral and Maxillofacial Surgery*, vol. 75, no. 7, pp. 1530–1541, 2017.

[5] T. Bressmann, H. Jacobs, J. Quintero, and J. C. Irish, "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*, vol. 33, no. 4, p. 204, 2009.

[6] A. Sinharay, D. Ghosh, P. Deshpande, S. Alam, R. Banerjee, and A. Pal, "Smartphone based digital stethoscope for connected health–a direct acoustic coupling technique," in *2016 IEEE First International Conference on Connected Health: Applications, Systems and Engineering Technologies (CHASE)*. IEEE, 2016, pp. 193–198.

[7] B. Raj, T. Virtanen, and R. Singh, *Techniques for Noise Robustness in Automatic Speech Recognition*.

[8] S. Gururangan, S. Swayamdipta, O. Levy, R. Schwartz, S. Bowman, and N. A. Smith, "Annotation artifacts in natural language inference data," in *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*. New Orleans, Louisiana: Association for Computational Linguistics, Jun. 2018, pp. 107–112. [Online]. Available: https://aclanthology.org/N18-2017

[9] M. Parmar, S. Mishra, M. Geva, and C. Baral, "Don't blame the annotator: Bias already starts in the annotation instructions," 2022. [Online]. Available: https://arxiv.org/abs/2205.00415

[10] A. Koenecke, A. Nam, E. Lake, J. Nudell, M. Quartey, Z. Mengesha, C. Toups, J. R. Rickford, D. Jurafsky, and S. Goel, "Racial disparities in automated speech recognition," *Proceedings of the National Academy of Sciences*, vol. 117, no. 14, pp. 7684–7689, 2020.

[11] A. Caliskan, J. J. Bryson, and A. Narayanan, "Semantics derived automatically from language corpora contain human-like biases," *Science*, vol. 356, no. 6334, pp. 183–186, 2017.

[12] Y. Zhang, Y. Zhang, B. M. Halpern, T. Patel, and O. Scharenborg, "Mitigating bias against non-native accents," *Submitted to Interspeech 2022*.

[13] B. Vachhani, C. Bhat, and S. K. Kopparapu, "Data augmentation using healthy speech for dysarthric speech recognition." in *Interspeech*, 2018, pp. 471–475.

[14] C. Espana-Bonet and J. A. Fonollosa, "Automatic speech recognition with deep neural networks for impaired speech," in *International Conference on Advances in Speech and Language Technologies for Iberian Languages*. Springer, 2016, pp. 97–107.

[15] S. Hahm, D. Heitzman, and J. Wang, "Recognizing dysarthric speech due to amyotrophic lateral sclerosis with across-speaker articulatory normalization," in *Proceedings of SLPAT 2015: 6th Workshop on Speech and Language Processing for Assistive Technologies*, 2015, pp. 47–54.

[16] Y. Qin, T. Lee, S. Feng, and A. P.-H. Kong, "Automatic speech assessment for people with aphasia using tdnn-blstm with multi-task learning." in *Interspeech*, 2018, pp. 3418–3422.

[17] M. M. Hakkesteegt, M. P. Brocaar, and M. H. Wieringa, "The applicability of the dysphonia severity index and the voice handicap index in evaluating effects of voice therapy and phonosurgery," *Journal of Voice*, vol. 24, no. 2, pp. 199–205, 2010.

[18] J. Oates, "Auditory-perceptual evaluation of disordered voice quality," *Folia Phoniatrica et Logopaedica*, vol. 61, no. 1, pp. 49–56, 2009.

[19] M. Windrich, A. Maier, R. Kohler, E. Nöth, E. Nkenke, U. Eysholdt, and M. Schuster, "Automatic Quantification of Speech Intelligibility of Adults with Oral Squamous Cell Carcinoma," *Folia Phoniatrica et Logopaedica*, vol. 60, no. 3, pp. 151–6, 04 2008.

[20] H. Kim and S. Nanney, "Familiarization effects on word intelligibility in dysarthric speech," *Folia Phoniatrica et Logopaedica*, vol. 66, no. 6, pp. 258–264, 2014.

[21] H. Kim, "Familiarization effects on consonant intelligibility in dysarthric speech," *Folia Phoniatrica et Logopaedica*, vol. 67, no. 5, pp. 245–252, 2015.

[22] T. Kaneko and H. Kameoka, "Cyclegan-vc: Non-parallel voice conversion using cycle-consistent adversarial networks," in *2018 26th European Signal Processing Conference (EUSIPCO)*. IEEE, 2018, pp. 2100–2104.

[23] B. M. Halpern, R. van Son, M. van den Brekel, and O. Scharenborg, "Detecting and analysing spontaneous oral cancer speech in the wild," *Proc. Interspeech 2020*, pp. 4826–4830, 2020.

[24] L. F. Gallardo, S. Möller, and M. Wagner, "Importance of intelligible phonemes for human speaker recognition in different channel bandwidths," in *Proc. Interspeech 2015*, 2015, pp. 1047–1051.

[25] P. Janbakhshi, I. Kodrasi, and H. Bourlard, "Pathological speech intelligibility assessment based on the short-time objective intelligibility measure," in *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2019, pp. 6405–6409.

[26] X. Tian, J. Wang, H. Xu, E. S. Chng, and H. Li, "Average Modeling Approach to Voice Conversion with Non-Parallel Data." in *Odyssey*, vol. 2018, 2018, pp. 227–232.

[27] D. Van Lancker, J. Kreiman, and K. Emmorey, "Familiar voice recognition: Patterns and parameters part i: Recognition of backward voices," *Journal of phonetics*, vol. 13, no. 1, pp. 19–38, 1985.

LIST OF PUBLICATIONS

16. T. Tienkamp, R. van Son, and B. M. Halpern, "Objective speech outcomes after surgical treatment for oral cancer: An acoustic analysis of a spontaneous speech corpus containing 32.850 tokens," *Submitted to Journal of Communication Disorders*

15. B. M. Halpern, S. Feng, R. van Son, M. van den Brekel, and O. Scharenborg, "Automatic evaluation of spontaneous oral cancer speech using ratings from naive listeners," *Submitted to Speech Communication*

14. ——, "Low-resource automatic speech recognition and error analyses of oral cancer speech," *Speech Communication*, vol. 141, pp. 14–27, 2022

13. W.-C. Huang, B. M. Halpern, L. P. Violeta, O. Scharenborg, and T. Toda, "Towards identity preserving normal to dysarthric voice conversion," in *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2022, pp. 6672–6676

12. Y. Zhang, Y. Zhang, B. M. Halpern, T. Patel, and O. Scharenborg, "Mitigating bias against non-native accents," *Accepted at Interspeech*, 2022

11. B. M. Halpern, T. Rebernik, T. Tienkamp, R. van Son, M. v. d. Brekel, M. Wieling, M. Witjes, and O. Scharenborg, "Manipulation of oral cancer speech using neural articulatory synthesis," *Submitted to Interspeech*, 2022

10. L. Prananta, B. M. Halpern, S. Feng, and O. Scharenborg, "The effectiveness of time stretching for enhancing dysarthric speech for improved dysarthric speech recognition," *Accepted at Interspeech*, 2022

9. B. M. Halpern, J. Fritsch, E. Hermann, R. van Son, O. Scharenborg, and M. Magimai-Doss, "An objective evaluation framework for pathological speech synthesis," in *Speech Communication; 14th ITG Conference*. VDE, pp. 1–5

8. M. Illa, B. M. Halpern, R. van Son, L. Moro-Velázquez, and O. Scharenborg, "Pathological voice adaptation with autoencoder-based voice conversion," in *11th ISCA Speech Synthesis Workshop*. ISCA, 2021, pp. 19–24

7. S. Feng, B. M. Halpern, O. Kudina, and O. Scharenborg, "Towards inclusive automatic speech recognition," *Submitted to Computer Speech and Language. Preprint with "Quantifying Bias in Automatic Speech Recognition" is available on arXiv.*

6. B. M. Halpern, F. Kelly, and A. Alexander, "Speaker-informed speech enhancement and separation," *IAFPA*, 2021

5. B. M. Halpern, R. van Son, M. van den Brekel, and O. Scharenborg, "Detecting and Analysing Spontaneous Oral Cancer Speech in the Wild," in *Proc. Interspeech 2020*, 2020, pp. 4826–4830

4. B. Halpern, F. Kelly, R. van Son, and A. Alexander, "Residual Networks for Resisting Noise: Analysis of an Embeddings-based Spoofing Countermeasure," in *Proc. The Speaker and Language Recognition Workshop (Odyssey 2020)*, 2020, pp. 326–332

3. B. Halpern and F. Kelly, "Can deepfake voices steal high-profile identities?" *IAFPA*, 2022

2. T. Tienkamp, T. Rebernik, B. Halpern, R. van Son, S. A. de Viscscher, M. Witjes, and M. Wieling, "Quantifying changes in articulatory working space following oral cancer treatment using electromagnetic articulography," *Speech Motor Control*, 2022

1. T. Rebernik, B. Halpern, T. Tienkamp, R. Jonkers, A. Noiray, R. van Son, M. van den Brekel, M. Witjes, and M. Wieling, "The effect of masking noise on oral cancer speech acoustics and kinematics," *Speech Motor Control*, 2022

ACKNOWLEDGEMENTS

I would like to thank my family. My mother sent me to programming classes when I was 10, which gave me an early foundation in programming that was tremendously helpful in this process. The work discipline that I learnt from my father was essential in the last four years. Lastly from my family, I would like to thank my sisters for introspective discussions, which taught me how should I approach this discipline.

I would like to thank Leo for his continued support during my life. Your death shocked me. It makes me especially sad that you never knew what I was working on, yet, you seem to have suffered so much from my absence.

I would like to thank Robert Toth for his continued support and availability in the last eight years. He is a great friend and a great scientist. Your advice was important when I was doubtful about my own scientific standards (imposter syndrome) and also at times when I was doubtful about others' scientific standards (which we coined reverse imposter syndrome).

I would like to express my gratitude to my supervisors. Most importantly, I would like to thank Odette Scharenborg who volunteered to be my co-promoter after a year of hopeless pathfinding. Odette made my PhD into a more interesting journey than it would have ever been without her. You taught me scientific writing, and your uplifting, enthusiastic, and analytic approach to research made me a better researcher. I feel privileged to have worked with you.

I would like to thank Rob van Son for all the advice and interesting discussions we had with each other. Whether I wanted to discuss the pointlessness of machine learning, the existence of phonemes, or ill-posed matrix inversion, you always had a good take on it. Furthermore, I would like to thank Michiel. Your kindness, punctuality and flexibility supported my research even at the most challenging times.

During my PhD, I spent a lot of time with other research groups, which I am grateful for. In Switzerland, Mathew taught me how an idea sometimes just has to grow. I appreciate the discussion I have had with everyone there, especially Julian, who had a sobering sense of cynical honesty, which should be a more common thing in academia.

I would really like to thank people from Oxford Wave Research, and how their ideas cross-pollinated my research. Their business perspective made me a better organiser and taught me that being a good engineer is as important as being a good researcher. This value is important to my integrity, standards, and how I approach other people's research. I would like to thank Anil for important discussions on intermittent fasting and calorie burning, Oscar for venting about Kaldi (and how horrible it can be), and Finnian

for his great ideas and for helping my writing. I would like to thank all other people at OWR, whom I had any interaction with: Ekrem, Jithin, Linda, Nikki, Sam, Tom, TC and Zofia.

I would like to thank people from the Multimedia Computing Lab. Especially the members of our "Monday Mental Health Club" - Omar, Alberto, Sandy, Manel, and Jay. But also thanks to Li, Xiuxiu, and others I have had any interactions with. I would also like to thank current and past members of the SALT group (44 people at the time of writing!). Most importantly, I should thank Siyuan – all the interactions and help that I had from you improved my professional skills a lot. I really regret that due to the pandemic we could meet so little physically. Thank you also Tanvina for providing me professional and swift help even in times when you were under the tightest of deadlines.

My master students - Anouck, Janay, Karl, Kirsten, Luke, Marc, Marjolein, Mathilde and Thomas – thank you! All of you provided an interesting lesson to me, not just professionally! Seeing some of your growth and development during the process made me understand my own shortcomings more deeply.

I would like to thank Ruud de Jong and all the IT support staff at TU Delft. I firmly believe that good technical support staff is essential for quality research.

People from the Groningen Speech Lab – I would also like to thank you for all the support I received from you, and the great times we had, including inviting me to the retreat. Teja – thank you for listening to my constant "data anxiety" and encouraging me when I had to stand up for myself. Thomas – thank you for putting up with my often unprofessional behaviour, and helping me out in everything. Bartelds – thank you for the football match and all the interesting discussions that we had during the retreat.

Thanks to all the OOA students at the NKI – most importantly Kilian for helping me with settling into the Netherlands at a critical time. I would like to thank Marise too, not just for the cookies (haha), but for her honesty, and introspective discussions on work-life balance. I still do not understand how can you eat pepernoten in a sandwich, though.

Lastly, there are some people whom I do not wish to name here but who supported my research in some way or another. I want to thank you too.