university of
groningen

campus fryslân

Spraakherkenning, wa is da? —

Bias in Flemish Speech Recognition

Aaricia Herygers
Master of Science in Voice Technology

Supervisor:	Dr. Vass Verkhodanova
Second Reader:	Dr. Matt Coler
External Advisor:	Dr. Odette Scharenborg
External Advisor:	Prof. Dr. Munir Georges

Academic year: 2021—2022

Contents

Preface **3**

Abstract **4**

1 Introduction **5**

2 Background **7**

2.1 Introduction to ASR . . . 7

2.2 Relevant ASR Models . . . 9

2.3 Motivation for Flemish ASR . . . 13

2.4 Bias in ASR . . . 14

2.4.1 Regional Bias . . . 15

2.4.2 Gender Bias . . . 16

2.4.3 Age Bias . . . 18

2.4.4 Non-Nativeness Bias . . . 20

3 Method **21**

3.1 Corpora . . . 21

3.1.1 Training . . . 21

3.1.2 Testing . . . 22

3.2 ASR Model . . . 24

3.3 Error Analysis . . . 24

4 Results **25**

4.1 Acoustic Model Results . . . 25

4.1.1 Regional and Age Bias . . . 25

4.1.2 Gender Bias . . . 26

4.1.3 Non-Nativeness Bias . . . 26

4.2 Results with RRNLM Rescoring . . . 27

4.2.1 Regional and Age Bias . . . 27

4.2.2 Gender Bias . . . 27

4.2.3 Non-Nativeness Bias . . . 28

4.3 Phoneme Error Analysis . . . 29

5 Discussion and Conclusion **30**

6 References **32**

Preface

While finishing the formatting of this thesis past midnight, I still want to take the time to thank some people. Thank you Vass and Matt for providing quick feedback and responses throughout the year and especially my last-minute thesis writing period. Thank you Odette for introducing me to other bias researchers and supporting me. Thank you Munir for your swift feedback and for taking a chance on me. Thank you Tanvina for all your help on getting the code to run, even when the HPC let us down. Thank you Siyuan and Bence for your last-minute tips. Thank you Mariano for your help and support. Thank you Kevin for wanting to tackle the code labyrinth with me. Thank you Laufje for always being there.

Abstract

Sociolinguistic factors such as age (Vipperla, Renals, & Frankel, 2008) and gender (Tatman, 2017) have been shown to impact the performance of various automatic speech recognition (ASR) models. Previous research has touched upon such performance discrepancies, uncovering biases in ASR models, but has often focused on the English language (e.g., Kathania, Reddy Kadiri, Alku, & Kurimo, 2020; Tatman & Kasten, 2017; Vipperla, Renals, & Frankel, 2010). However, as these systems are used worldwide, finding biases in different languages is of high importance. With this thesis, I extend recent research by Feng, Kudina, Halpern, and Scharenborg (2021), who sought to find biases based on age, region, gender, and non-nativeness in a Dutch ASR model. Like Feng et al. (2021), I use the Netherlandic Dutch data from the Spoken Dutch Corpus (Oostdijk, 2000) to train a hybrid deep neural network-hidden Markov model (DNN-HMM). However, the previous study did not take into account the various regional variants of Belgian Dutch, which is also known as Flemish, e.g., West Flemish, and Brabantian (Odijk, 2012). I therefore evaluate the model using the Flemish data from the JASMIN-CGN corpus (Cucchiarini, Van hamme, van Herwijnen, & Smits, 2006). The evaluation confirms a bias against speakers from West Flanders and Limburg, as well as against children, male speakers, and non-native speakers. In addition, the discussion of the findings includes an analysis of the most misrecognized phonemes. The current study contributes to a better understanding of bias, and subsequently inclusivity, in ASR.

1 Introduction

Technology is ubiquitous—the omnipresence of the Internet of Things came with a rise in voice-automated devices, such as smart speakers. In fact, a survey showed that in 2019 one in four Americans owned such a device (Auxier, 2019) and a 2022 report demonstrated that currently every other American owns a smart home device, with one third of them using voice assistants (Kinsella, 2022). However, these devices do not perform equally well for everyone trying to use their speech to control the room temperature, create a shopping list, call their family, or even order their favorite drink from Starbucks (Hoy, 2018)[1].

The accuracy with which the voice commands are recognized by the automatic speech recognition (ASR) system is dependent on various factors. For example, ASR systems experience difficulties in distinguishing commands when multiple people are speaking (Qian, Weng, Chang, Wang, & Yu, 2018), when the environment is noisy (Hamidi, Satori, Zealouk, & Satori, 2020; Kathania, Kadiri, Alku, & Kurimo, 2021), or when the speech contains multiple languages (Sreeram & Sinha, 2020). Another influence on the accuracy of ASR are the speakers themselves (Benzeghiba et al., 2007).

In essence, depending on the speech data that was used to train the system, the acoustic model specifically, speech by certain people, such as the elderly (Vipperla et al., 2008; 2010), tends to be recognized more accurately than speech by others. This difference in accuracy between various (groups of) people is called *bias*. Bias has already been the topic of many studies in adjacent fields such as machine learning (e.g., Hellström, Dignum, & Bensch, 2020; Mehrabi, Morstatter, Saxena, Lerman, & Galstyan, 2021), natural language processing (e.g., Blodgett, Barocas, Daumé, & Wallach, 2020; Sun et al., 2019), chatbot design (e.g., Feine, Gnewuch, Morana, & Maedche, 2019; Zabel & Otto, 2021), and researchers have recently started to investigate bias in automatic speaker recognition (Toussaint Hutiri & Ding, 2022). The contemporary area of research into bias in ASR seeks to uncover sociolinguistic factors that negatively impact the recognition accuracy of ASR systems, and where possible, mitigate the bias (e.g., Peri, Somandepalli, & Narayanan, 2022; Zhang, 2022; Zhang, 2022).

A study conducted by Feng et al. (2021) forms the basis of this thesis; they quantified bias in a Dutch ASR model and found that it performed worse for certain speaker groups, i.e., had a higher word error rate (WER) compared to other groups. In other words, differences in performance were found between speakers of different ages and genders, and from different regions. Furthermore, the model performance differed between native speakers and non-native speakers of Dutch. The study also showed that the model did not perform well on speech from Flanders, the northern, Dutch-speaking part of Belgium.

In light of the regional varieties within Flemish speech (Odijk, 2012) and the differences in their intelligibility (Impe & Geeraerts, 2008; Van Bezooijen & Van den Berg, 1999a; 1999b), this thesis seeks to answer the following questions:

RQ 1: Using Flemish data, does a Dutch hybrid deep neural network-hidden Markov model ASR system with language model rescoring exhibit bias in terms

[1] I want to provide an anecdote by an acquaintance as an example; recently they got a TV remote with a speech recognition button, which bypasses the need to browse the digital program guide manually. Usually, it will recognize the command "Go to Netflix". Actually, even the command "Netflix" will suffice. However, after this acquaintance had taken out their denture, the command sounded more like "Njetfwlisxsch" (/ɲɛtfwɬɪskʃ/), to which the synthetic voice replied "Please repeat that". After repeating "Njetfwlisxsch, dwamn it", the synthetic voice replied "Please do not curse".

of a difference in WER when comparing speakers from different regions, in different age groups, with different genders, and native versus non-native speakers?

RQ 2: Compared to the acoustic model, does the bias decrease when using language model rescoring?

The hypotheses are as follows:

H1 Region: I expect that there will be a difference in WER when comparing regions, following the results from Impe and Geeraerts (2008) and Van Bezooijen and Van den Berg (1999a; 1999b).

H1.1 More specifically, I expect a higher WER for West Flemish speakers compared to those from the transitional and central regions

H1.2 Additionally, I expect a higher WER for Limburgish speakers compared to those from the transitional and central regions.

H2 Age: I expect that there will be a difference in WER when comparing age groups, following the results from Feng et al. (2021).

H2.1 More specifically, I expect a higher WER for children than for youngsters.

H2.2 Additionally, I expect a higher WER for elderly speakers than for youngsters.

H3 Gender: I expect that there will be a higher WER for male speakers than for female speakers, following the results from Feng et al. (2021), because this study makes use of the same corpora (Spoken Dutch Corpus (Oostdijk, 2000), and Jasmin-CGN (Cucchiarini et al., 2006)).

H4 Non-nativeness: I expect that there will be a higher WER for non-native speakers than for native speakers, following the results from both Feng et al. (2021) and Van Wijngaarden (2001).

If H1 through H4 are all rejected, that would suggest that the model recognizes all types of speech well or that the biases lie somewhere else, which future research could discover. This would indicate that the specific training and test data used are of high importance. In any case, the findings of this research will provide further insights into bias in ASR.

In the following sections, I will introduce the topic of automatic speech recognition (Section 2.1) and describe relevant ASR models (Section 2.2). In Section 2.3, I will describe the motivation for a study on Flemish ASR. Then I will provide an overview of the literature on different biases (Section 2.4), followed by a description of the used corpora in Section 3.1, ASR model in Section 3.2, and phoneme error analysis in Section 3.3. After that, I will present the bias results as well as the results from the error analysis (Section 4). I will conclude this thesis by providing a detailed account of the implications of the research (Section 5).

It is important to note that most of the literature makes reference to 'gender' bias even though often the speaker data contains information that pertains to binary labels referring to 'sex', making the term 'gender' inaccurate. Similarly, the notion of a 'native speaker' is considered "problematic" (Cheng et al., 2021), though often used in previous literature. These topics, however, are beyond the scope of this thesis. Hence, I will follow the conventions of the literature, and thus use the terms 'gender' and '(non-) native speaker', for uniformity.

2 Background

This section first introduces the reader to ASR by describing the history of ASR systems and their basic components. Section 2.2 then provides an overview of various relevant ASR models. In Section 2.3 I outline the Flemish context, stating why this research is necessary. The different types of biases researched in this study are presented in Section 2.4: bias due to regional language varieties (Section 2.4.1), gender (Section 2.4.2), age (Section 2.4.3), and non-nativeness (Section 2.4.4).

2.1 Introduction to ASR

The main means of human communication is speech. Since the 1950s, efforts have been made to extend this mode of communication to computers, bridging the gap between multiple fields, including signal processing, pattern recognition, linguistics, and more (Rabiner & Juang, 1993). In essence, ASR, also known as speech-to-text, converts spoken language into written language. Science fiction has implemented speech recognition multiple times, for instance in 1966, a universal translator that made use of ASR was featured in *Star Trek*[2]. Another example is *Battlestar Galactica* from 1978, in which Commander Adama keeps a journal by dictating his stories to a speech-to-text machine. However, ASR is not merely science fiction.

ASR started with Audrey, the Automatic Digit Recognition machine created at Bell Labs, which could accurately recognize digits between zero and nine, as spoken by its creator (Davis, Biddulph, & Balashek, 1952). A decade later, IBM added six words to their Shoebox recognizer, which made it possible for the machine to perform mathematical calculations: minus, plus, subtotal, total, false, and off (IBM, n.d.-a; Sikka, n.d.). In 1976, the Harpy system was developed at Carnegie Mellon University (Lowerre, 1976). It was already able to recognize over one thousand words. Further significant systems include IBM's Tangora (IBM, n.d.-b) in the mid-1980s, which sported a vocabulary of twenty thousand words and made use of a statistical method: the Hidden Markov Model (HMM), which is still in use to this day (O'Shaughnessy, 2019) (see Section 2.2 for an explanation of HMMs). Shortly before the new millennium Dragon Systems developed Dragon Dictate, where every word had to be followed by a pause, and later Dragon NaturallySpeaking (Zumalt, 2005). A more detailed overview of the advancements (in statistical modeling) in speech and speaker recognition is available in Furui (2005; 2010).

The following decades saw fast-paced developments with Microsoft implementing dictation in Office in 2002 (Zumalt, 2005) and Google providing voice-automated business search in 2006 (Bacchiani, Beaufays, Schalkwyk, Schuster, & Strope, 2008) as well as voice input for Google Maps and Google Search in 2008 (Schalkwyk et al., 2010). After their influential Shoebox and Tangora, in 2011 IBM's Watson successfully competed in the TV show *Jeopardy!*, showing the impact of and opportunities for ASR and natural language processing. Apple's Siri integration into iOS followed that same year (Silver, 2022), paving the way for further voice assistants.

[2] A short clip of the universal translator, or rather its malfunction, from *Star Trek: Discovery* can be found here: `https://youtu.be/lIKppXg0sYw`

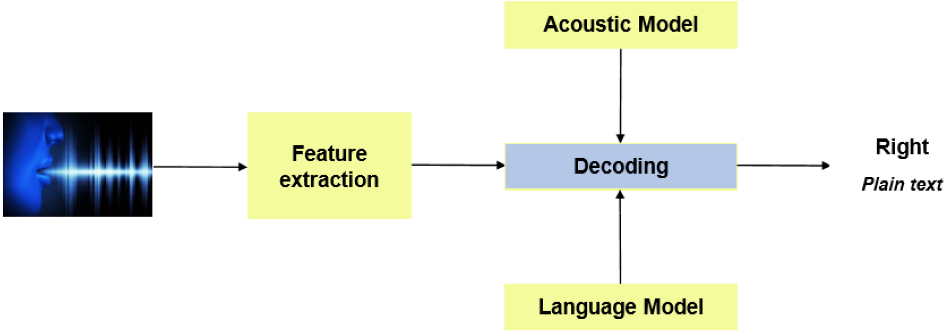

Figure 1: The basic components of an ASR system (Pervaiz et al., 2020)

Quickly thereafter followed Google Now, Microsoft's Cortana, Amazon's Alexa, and Google Home, shaping the ASR landscape.

The basic components of most ASR systems are presented in Figure 1 by Pervaiz et al. (2020: p. 3). Speech, which is a continuous signal, is converted into a digital signal, "a pattern of bits" (Priya & Kannamal, 2020: p. 3). Through sampling, the continuous signal is converted into a discrete one, making it storable in computers. From this signal, background noise is disregarded and a range of features is extracted as vectors. Among these features, Mel-Frequency Cepstral Coefficients (MFCCs) are commonly used—this study makes use of them as well (see Section 3.2 for a description of the ASR model). MFCCs are derived by following a number of steps such as taking the Fourier transform of a signal, and are based on the mel scale (Stevens, Volkmann, & Newman, 1937); a scale that, contrary to the Hertz scale, is non-linear and thus better reflects human pitch perception. This makes them good candidates for feature extraction (Kamath, Liu, & Whitaker, 2019). The extracted vectors are then given as input to the decoder, which consists of an acoustic model (AM) and language model (LM). The acoustic model, e.g., HMM, maps sound units to phonemes. A pronunciation lexicon will then attempt a mapping of the phoneme to a grapheme, e.g., the Dutch voiced velar fricative /ɣ/ to the grapheme <g>. After combining the sounds into a sequence, ideally a word, the LM calculates the likelihood of that word appearing in the given context (sentence), taking into account the language constraints (syntax and semantics) (O'Shaughnessy, 2019). For example, the phrase "Eye sea a duck" would be less likely than "I see a duck". The system subsequently outputs the text with the highest likelihood.

To evaluate how well an ASR system works, one can calculate the word error rate (WER). It is "the most commonly used metric for speech recognition" (Kamath et al., 2019: p. 387) and is calculated as follows:

$$WER = \frac{(S + I + D)}{N} * 100$$

Where:

- S is the number of substitutions at word level.
- I is the number of insertions at word level.
- D is the number of deletions at word level.
- N is the total number of words in the reference.

2.2 Relevant ASR Models

This section introduces some of the different models that are currently in use and are relevant for the understanding of the ASR model used in this thesis: hidden Markov models (HMMs), deep neural networks (DNNs), time-delay neural networks (TDNNs), recurrent neural networks (RNNs), and bidirectional long short-term memory (BLSTM) models.

The 1970s and even more so the 1980s, with IBMs Tangora (IBM, n.d.-b), brought an approach to speech recognition that is still applied today: the Hidden Markov Model. An HMM is a process in which a stochastic (i.e., random) sequence of observations is produced based on another hidden (i.e., presumed) stochastic process (Furui, 2005; 2010; O'Shaughnessy, 2019; Rabiner & Juang, 1986; 1993). Rabiner and Juang (1986) explain this concept through the example of a coin toss. The results of the coin toss are shared (i.e., heads or tails), but the tossing is not visible. In ASR, the inputs for HMMs are the extracted feature vectors, which come from the speech signal. These are observations based on the internal speech production (e.g., tongue position), which is hidden (O'Shaughnessy, 2000). Figure 2 shows a basic HMM with three variables, or states, that model the <u> in the word "cup" (Kumar, 2018). The nodes represent the states and the arrows represent the state transitions, i.e., transitioning from one state to another (horizontal arrows) or staying in the same state (curved arrows). Figure 3 shows the concatenation of the HMMs of <c>, <u>, and <p> (Kumar, 2018). The probability of the state transitions in an HMM were often modeled using Gaussian mixture models (GMMs), creating a GMM-HMM. A GMM consists of multiple Gaussians, which are probabilistic (Contreras Carrasco, 2019).

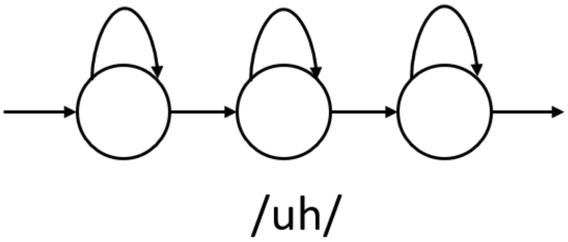

Figure 2: Basic HMM for <u> in "cup" with three states, each containing a forward transition and a self-loop (Kumar, 2018)

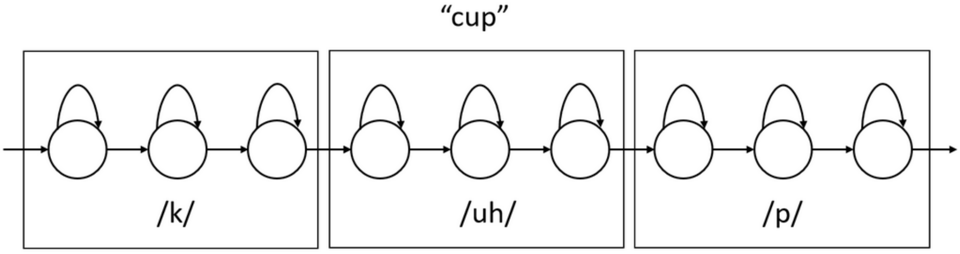

Figure 3: Concatenated HMM for "cup" with three states for each phoneme, each containing a forward transition and a self-loop (Kumar, 2018)

However, in the late 1980s GMM-HMMs started to be, albeit slightly, outperformed by hybrid models where the state transitions were modeled through artificial neural networks (ANNs)

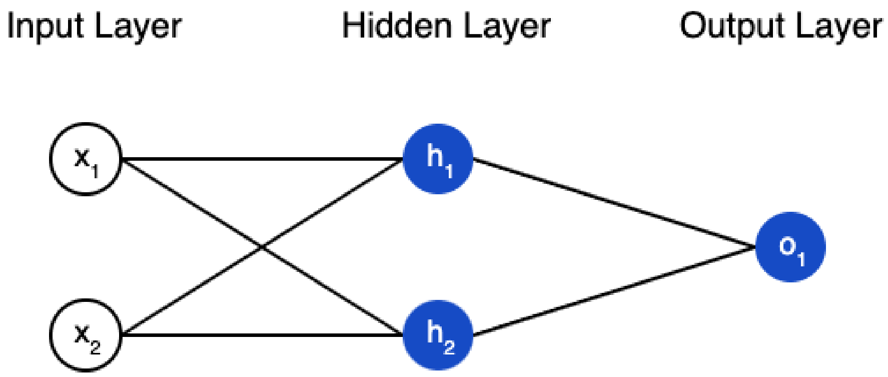

Figure 4: A basic neural network with two inputs, one hidden layer with two nodes, and one output node (Zhou, 2019)

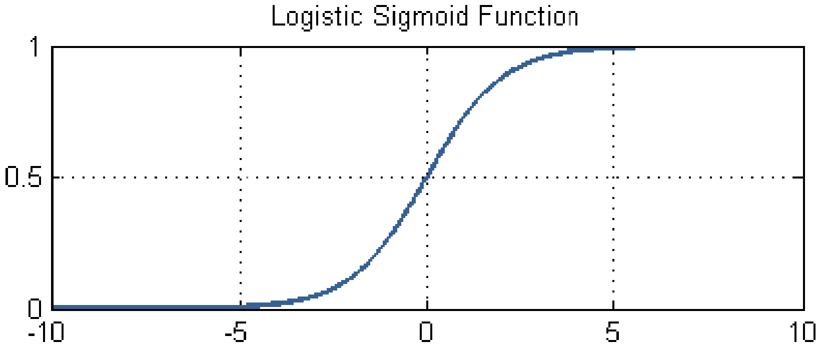

Figure 5: A sigmoid function (Chang et al., 2012)

(Dahl, Yu, Deng, & Acero, 2012). An ANN is built up of layers of nodes, also called *neurons* or *perceptrons*. Each neuron is given an input and a mathematical function to perform and then outputs the result. Figure 4 (Zhou, 2019) shows a basic ANN with two inputs (x_1, x_2) which are passed onto two neurons in the hidden layer (h_1, h_2). Following that, the output from the hidden layer is given to one neuron in the output layer (o_1). The mathematical calculation involves adding weights to the inputs and transforming this sum through an activation function. One such activation function is the sigmoid, or logistic, function (Mitchell, 1997), which is defined as:

$$\sigma(y) = \frac{1}{1 + e^{-}y}$$

The output of a sigmoid function lies between 0 and 1, as presented in Figure 5. Therefore, sigmoid functions are commonly used to predict a probability, such as the probability that the phrase “Eye sea a duck” would occur.

As computer power grew over the years, it became possible to add more intermediate hidden layers to ANNs, which allow for “greater flexibility in [machine learning]” (O’Shaughnessy, 2019: p. 3467). When a neural network entails many hidden layers, it is called a deep neural network (DNN). A combination of such a DNN with an HMM was proposed by Dahl et al. (2012). Figure 6 shows a diagram of their hybrid architecture in which the state transitions are modeled by the HMM and the triphone observation likelihoods are modeled by the DNN. The nodes S_1 through S_K depict the states. The transition probabilities for both forward transitions and self-loops are given by $a_{S_1S_1}$ through $a_{S_KS_K}$. The observation probabilities from the HMM are then given to the DNN, in which $h^{(1)}$ through $h^{(M)}$ are the hidden layers, v is the visible layer, and W_1 through W_M are the weights. The input for the visible layer is the speech signal.

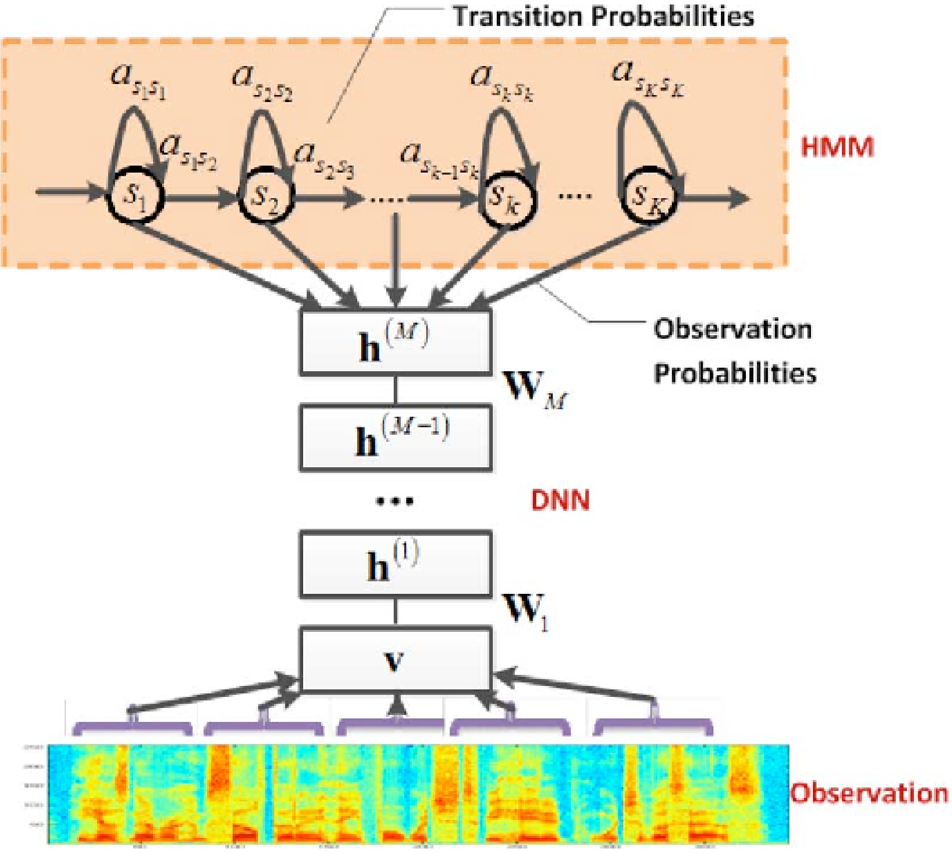

Figure 6: Hybrid DNN-HMM architecture by Dahl et al. (2012)

Despite their relative success, neural networks still displayed shortcomings in handling the time information of acoustic events. Hence, Waibel, Hanazawa, Hinton, Shikano, and Lang (1989) proposed a new architecture for phoneme recognition, specifically /b/, /d/, and /g/: time-delay neural network (TDNN). It integrates delays into its units and passes the sums through the sigmoid function, which was opted for "due to its convenient mathematical properties" (p. 329). Therefore, it is possible for TDNNs to "represent temporal relationships between acoustic events" (p.329), while still ensuring time invariance. The TDNN architecture is provided in Figure 7. Specifically, it depicts the 16 coefficients for the speech input "BA". Black squares represent positive values and gray squares represent negative values. The input layer is fully connected to the first hidden layer, which contains eight time-delay units. A second hidden layer contains three such units. By integrating over the values given by the second hidden layer and passing this information to the output layer, the result (in this case /b/) is given.

In ASR, it is desirable and important to include contextual information over longer periods of time. While HMMs could model time predictions, the number of previous states is fixed. A more flexible solution is the recurrent neural network (RNN), as they have the ability to "[accumulate] information from each time step" (Kamath et al., 2019: p. 315). Therefore, the entire sequence is taken into account when making predictions. However, during model training, RNNs may suffer from a problem known as a *vanishing* or *exploding gradient* (for a more detailed

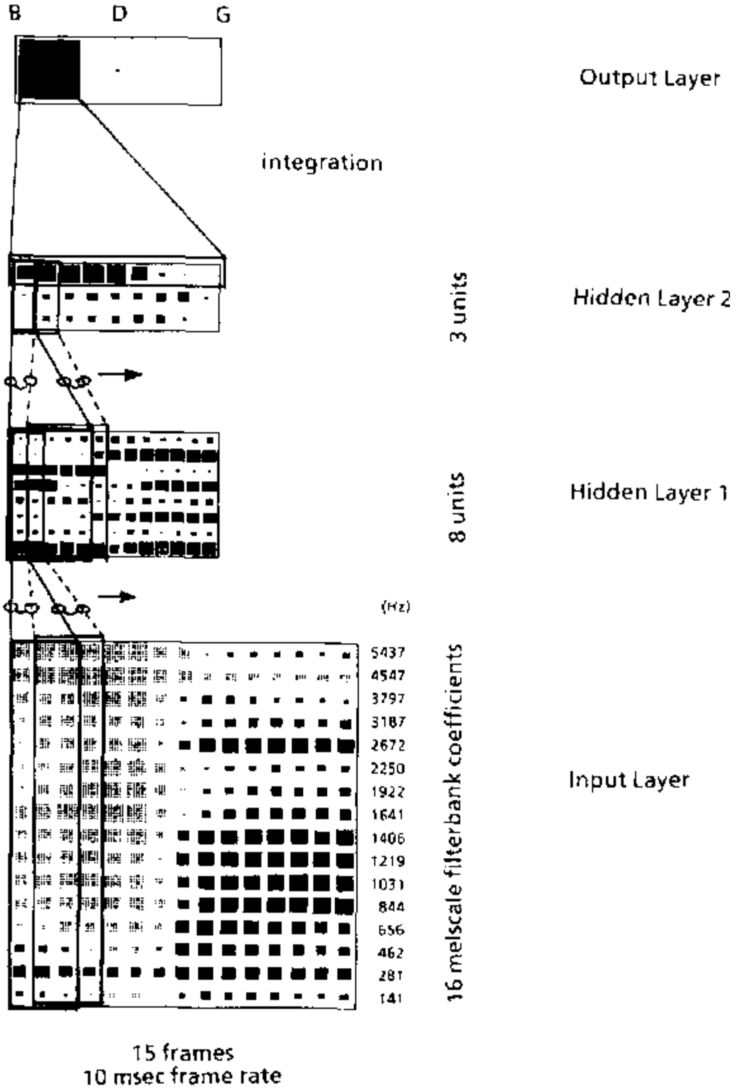

Figure 7: TDNN architecture by Waibel et al. (1989)

account on training a model, i.e., forward and backward propagation, please refer to Kamath et al. (2019)). Optimization techniques such as regularization can reduce the chances of such a vanishing or exploding gradient occurring.

Another possible method to prevent gradient problems is the long short-term memory (LSTM) model (Hochreiter & Schmidhuber, 1997). LSTMs have *memory cells*, which are a type of hidden state that stores content and include three types of *gates*: input, output, and forget gates. These gates are mechanisms that address the problems in previous models; seeing as HMMs are only able to 'remember' a fixed number of states and RNNs can 'remember' too much, so that training may come to a halt, LSTMs are designed to "learn the conditions for when to forget, ignore, or keep information in the memory cell" (Kamath et al., 2019: p. 324). Combining an LSTM with another LSTM in the other direction, meaning that not only the previous, but also the following context is taken into account, results in a bidirectional LSTM (BLSTM). An example of such a bidirectional architecture is given in Figure 8 (Ji, Han, Hou, Song, & Du, 2020). It shows an input layer that is connected to a forward layer as well as to a backward layer. The output stems from the information of both the forward and backward layers. Multiple such structures together model the sequential information from a speech signal.

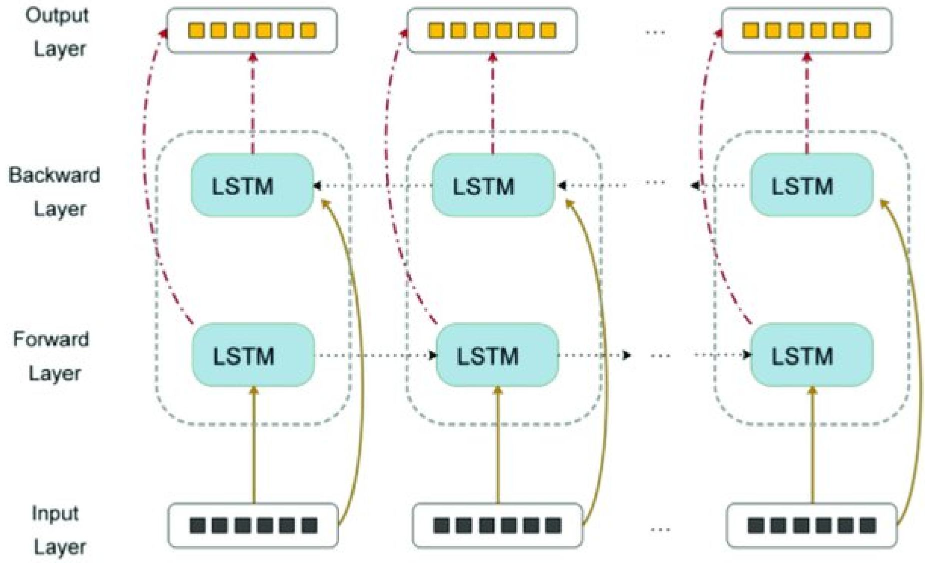

Figure 8: BLSTM architecture by Ji et al. (2020)

2.3 Motivation for Flemish ASR

This section provides the reader with an overview of the Flemish context. It first outlines the study on which this thesis is based, followed by an explanation on why this examination of Flemish is needed.

Feng et al. (2021) tested a Dutch "state-of-the-art" (p.1) speech recognizer on different kinds of speech. They quantified the bias with regard to various sociolinguistic factors: gender, age, region, and whether or not the speaker was a native speaker of Dutch. In addition, they compared the word error rates (WERs) of read speech and (semi-) spontaneous human-machine interaction (HMI). Speech of four regions in the Netherlands (classified as West, Transitional, North, and South) was tested. It was found that read speech was recognized more accurately than HMI in all cases. Additionally, the model exhibited bias against both native and non-native male speakers. The authors also point out that there is less variation in WERs across regions for read speech by children aged 7–11 and youngsters aged 12–16, which suggests that speech by adults over 60 years old has a larger *linguistic distance* (Geeraerts, Grondelaers, & Speelman, 1999; Impe, Geeraerts, & Speelman, 2008) from standard speech. Interregional differences in WER across the Netherlands can be noted as well: in HMI, speech from the South and North regions respectively show the highest and lowest WERs.

In addition to the investigation of bias in terms of WERs, the recognition errors were systematically analyzed in an in-depth examination of the phoneme error rate (PER). The report of this examination included the top 5 phonemes that occurred no less than 50 times. Here too, differences between the various speaker groups could be observed. It was found that the model performance, overall, was poor for vowels as spoken by non-native speakers, e.g., /œy/, /ʏ/, /y/, and /øː/. The authors attribute this to the level of difficulty to acquire these vowels for learners of Dutch. The reported misrecognition of phonemes from speech by native speaker in their respective age groups showed that children's /ʏ/, /h/, /ə/, /j/, youngsters' /ʏ/, /øː/, /h/, /ə/, and older adults' /h/, /x/, /ɛ/, /ə/ were poorly recognized.

In addition, the research featured speech from Flanders, which differs from Netherlandic Dutch on various levels such as lexical and phonetic (Impe et al., 2008; Steurs, Vandeghinste, & Daelemans, 2022). Across the board, the model is able to recognize Flemish the least accurately, specifically from Flemish children, with WERs of up to 66.4% for girls' HMI and 56.4% for boys' read speech. The analysis of the phoneme errors for Flemish showed that, overall, the phonemes /y/, /œy/, /ɑu/ were difficult to recognize.

However, the article does not fully capture possible biases in Flemish speech as the whole of Flanders was considered one region. Flemish as a variety of Dutch itself, though, has been increasingly studied (e.g., De Caluwe, Hüning, Vogl, & Moliner, 2012; Delarue & De Caluwe, 2015) and also consists of several language varieties (Odijk, 2012: West Flemish, East Flemish, Brabantian, Antwerpian, Limburgish) that are not all equally (mutually) intelligible.

For example, Van Bezooijen and Van den Berg (1999b) studied the objective intelligibility of three peripheral varieties of Dutch (Gronings, Limburgish, West Flemish) and Frisian, which is a separate language, as indicated by listeners from western and central provinces in the Netherlands. They found that the Gronings variety was the easiest to understand, whereas West Flemish and Frisian were equally difficult to understand. Another study by Van Bezooijen and Van den Berg from 1999a with both Dutch and Belgian listeners showed similar results. Although listeners from Belgium found Frisian the least intelligible, for Dutch listeners West Flemish was less intelligible than the varieties from Groningen, Limburg, and Friesland. This difference in intelligibility in human listeners may also indicate a difference in ASR performance.

Similarly, Impe and Geeraerts (2008) investigated the intelligibility of Standard Dutch compared to regional varieties. In their study, they asked students from West Flanders, Antwerp, Brabant, and Limburg to participate in a lexical decision task, various multiple choice questions, and a survey. They demonstrated that there is a significant effect of language varieties on the word intelligibility. More specifically, Standard Dutch and Brabantian words were recognized faster and with higher accuracy. In contrast, participants recognized the peripheral varieties West Flemish and Limburgish slower and with lower accuracy. While Antwerpian is often combined with Brabantian into a 'central' variety (e.g., Devos, 2006; Taeldeman, 2001), it was found to be significantly less intelligible than Brabantian, though more intelligible than the peripheral varieties. This shows that even within Flemish listeners there are differences in the intelligibility of regional varieties.

Consequently, as previously stated in H2, I hypothesize that these intelligibility differences will be noticeable in the comparison of ASR performance across Flemish varieties. Therefore, a study into the ASR of Flemish is required. However, seeing as most voice assistants have not included Belgian Dutch in their repertoire (Technologies, 2021)—in fact, Google only recently integrated a Flemish synthetic voice into their Google Assistant, in the fall of 2021 ("Philippe Geubels als Vlaamse stem van Google Assistent", 2021)—, I build upon Feng et al's study that included a Dutch model trained only on Netherlandic Dutch.

2.4 Bias in ASR

Bias in ASR is most commonly defined as the difference in recognition accuracy, or WER, between different groups of speakers. As the following subsections will demonstrate, the presence and scope of this bias is dependent on the data and ASR model that were used. Therefore, there is not one truth in bias research.

The following subsections provide a review of the literature on biases in ASR due to regional language variants, gender, age, and non-nativeness. In comparing these studies, I thus do not seek to invalidate previous research, but rather illustrate the various possible outcomes. However, it is clear that bias can present at various stages of a research workflow (Toussaint Hutiri & Ding, 2022).

2.4.1 Regional Bias

While speech recognition systems have become more prevalent only in recent years, studies into the recognition of regional varieties have been present, albeit relatively scarcely, for over two decades. Previous research has shown that speech in regional language varieties may increase the ASR model's WER as traditionally these models, more specifically the acoustic models and pronunciation lexica, are not developed using speech samples of non-standard language varieties (Lehr, Gorman, & Shafran, 2014). Using both Standard American English (SAE) and African American Vernacular English (AAVE) speech data, Lehr et al. (2014) compared the performance of a baseline GMM-based model in the IBM toolkit. They found an increase in WER of over 10% for AAVE compared to SAE. In contrast, their discriminative pronunciation model showed a 2% improvement for AAVE.

Similarly, German ASR models experience difficulty in recognizing within-country regional or non-standard speech. Thus, Schiel, Kipp, and Tillmann (1998) proposed to statistically model the pronunciation for each individual language variety with SAMPA units. However, having tested such modeling in the VERBMOBIL task, they found that it is only effective when "a *sufficient amount of reliably transcribed speech data*" is available (1998: p. 131, italics in original).

Following that, Beringer, Schiel, and Regel-Brietzmann (1998) considered the influence of "moderate regional variants" (p. 85) on the performance of the Daimler Benz HMM-based ASR system (Class, Kaltenmeier, & Regel-Brietzmann, 1993) by including dictionaries with regional pronunciations in the model training. However, no significant improvement over the baseline (30.91%) was found. They suspected that the ASR system might already capture the slight regional variation. Greater regional variation might lead to different results.

Baum, Erbach, Kommenda, Niklfeld, and Puig-Waldmüller (2001); Baum, Muhr, and Kubin (2001) reported the results of an evaluation study by Philips Speech Processing in which the performance of two acoustic models was compared: a model trained on German from Germany (GG) and a model trained on German from Austria (GA). It was found that, when testing with GA data, the GG model performed more poorly. In fact, a 7% increase in WER was found compared to the GA model: 23% vs. 16%. Later, even slightly adapting a model trained on GG to include Austrian pronunciations on a rule basis was found to have a positive impact on the recognition of Austrian speech (Wepner, 2021).

Comparable studies have been conducted for English. For example, Tatman (2017) investigated the automatic captions, based on Google's speech recognition system, for YouTube videos on the so-called "accent tag". In such videos, speakers often go through a word list which they pronounce in their regional language variety. Because the speakers pronounce isolated words, there is no context or sentence information available for the language model of the recognition system. A comparison of the captions' accuracy for speakers from five regions (Scotland, Georgia, New England, California, and New Zealand) demonstrated that they are less accurate for speakers speaking a Scottish variety of English relative to the other varieties.

In the same year, Tatman and Kasten (2017) compared WERs for automatic YouTube captions and Bing Speech transcriptions for recordings made in four "acoustically distinct" (p. 934) varieties of American English: General American, Northern Cities, Southern and Californian English. Although Xiong et al. (2017) presented a WER as low as 6.2% for Microsoft's 2016 conversational speech recognition system, the system behind Bing Speech, Tatman and Kasten found a mean WER of 45% in the Bing Speech transcriptions, though no significant differences across regional variety were found in this system. However, automatic YouTube captions, with a mean WER of 31%, did show such statistically significant differences. Contrary to the findings by Tatman (2017), where the lowest WER was found for Californian English, the Youtube ASR system performed the worst for the Californian variety, followed by the Southern and Northern varieties respectively, with the best performance reported for General American.

More recently, Nigmatulina, Kew, and Samardzic (2020) investigated non-standard variation in fourteen variants of Swiss German using the ArchiMob corpus (Samardžić, Scherrer, & Glaser, 2016). They still reported a WER of over 40% for a Swiss German TDNN model with dialectal transcriptions, and nearly 30% with normalized transcriptions. Better performance was found when using other types of error rates: 21.27% using a flexible WER (FlexWER), "which accounts for permissible spelling variants" (p. 21), and 14.64% using a character error rate. Interestingly, they noted that the amount of training data per variant did not considerably impact the recognition. For example, the lowest WER, approximately 20%, was found for the Graubünden variant, for which only one interview was included in the training data. Conversely, although the model was trained using twelve interviews in the Zurich variant, a WER of 40% was found. The authors indicate that (more) training data including a broader range of domains and a larger number of different speakers could improve the performance.

That being said, Dizdarevic, Hagmuller, Kubin, Pernkopf, and Baum (2004) introduced a regional variety recognizer for Austrian and German German using prosody information from the SpeechDat databases, which was then evaluated using the acoustic model for the respective variety, trained with the Hidden Markov Toolkit (HTK) (Young et al., 2000). Despite a fairly low variety recognition rate, fairly low WERs of 12.20% for Austrian German and 15.54% for German German were found.

A summary of the previously mentioned studies can be found in Table 1. For each publication, the evaluated (acoustic) model type is given, followed by the main result. The date and model type for the study by Philips Speech Processing were not provided.

2.4.2 Gender Bias

Against the backdrop of Doukhan, Poels, Rezgui, and Carrive's (2018) study, which uncovered gender inequality in French media, Garnerin, Rossato, and Besacier investigated the impact of media representation on the performance of ASR. For their investigation, they made use of a hybrid HMM-DNN with speaker adaptation in the Kaldi speech recognition toolkit (Povey et al., 2011) and a 5-gram LM using the SRILM toolkit (Stolcke, 2002). Looking at data from 1998 until 2003, Garnerin et al. found that two thirds of the speakers were male and male speakers occupied three quarters of the speaking time. Using this unequal data, they determined "that

Publication	Model	Result
Lehr et al. (2014)	GMM	Non-standard worse
	Discriminative pronunciation model	Improvement but still non-standard worse
Schiel et al. (1998)	HMM	Improvement when modeling pronunciation statistically
Beringer et al. (1998)	HMM	No significant improvements
Philips Speech Processing		Improvement when training on Austrian speech
Wepner (2021)	Rule-based	Improvement when including Austrian speech
Tatman (2017)	Youtube	Differences across varieties
Tatman and Kasten (2017)	Bing	No significant differences
	Youtube	Non-standard worse, differences across varieties
Nigmatulina et al. (2020)	TDNN	Swiss German worse
Dizdarevic et al. (2004)	HMM	German German worse

Table 1: Overview of publications on regional bias

gender is clearly a factor of variation in ASR performance" (2019: p. 8). In a 70-hour evaluation dataset, as demonstrated by WERs calculated at episode level for each speaker, gender impacted ASR performance. Unlike in the study by Feng et al. (2021), although there were differences in WERs across TV shows, on average, the model correctly recognized male speakers 24% better than female speakers.

Following those findings, (Garnerin, Rossato, & Besacier, 2020) proposed a reflection on ASR evaluation practices. In their subsequent study 2021, they conducted an experiment into female representation in the LibriSpeech corpus (Panayotov, Chen, Povey, & Khudanpur, 2015), which is a widely-used corpus of audiobook recordings, and recognition accuracy based on gender. The end-to-end model was trained three times with the ESPnet toolkit (Watanabe et al., 2018), using LibriSpeech data that included 30%, 50%, or 70% female speech. They concluded that increasing the amount of speech data by women decreased the WER, though not for the whole model, but rather only for female speakers.

In her 2017 work on automatic YouTube captions, Tatman not only investigated the effect of regional varieties on WER, but she also examined the influence of gender and found that female speakers were recognized less accurately than male speakers. However, Tatman and Kasten (2017) did not replicate these results with YouTube captions and Bing Speech transcriptions.

Moreover, like Feng et al. (2021), earlier research has also found the reverse: Sawalha and Abushariah (2013) examined the performance of an Arabic HMM-based ASR system trained on a Modern Standard Arabic speech corpus (Abushariah et al., 2012). With 5.29% and 7.48% as WERs for female and male speakers respectively, they revealed a bias against male speakers. Sawalha and Abushariah suggested that the cause of this difference in WER might be that the higher amplitudes and frequencies in female voices positively influenced the MFCCs and thus generated clearer feature vectors.

Adda-Decker and Lamel (2005) analyzed the performance of HMM-based ASR systems (Schwartz et al., 2004) for French and for American English using broadcast news and spontaneous telephone conversations. They found lower WERs for female speakers in both languages, i.e., 9.4% compared to 8.7% for French data and 12.0% compared to 9.4% for English data. Upon further analyses on the lexical and acoustic levels, they attributed the discrepancy to men's speech deviating more from standard pronunciations as well as including more disfluencies.

Table 2 offers a synopsis of the above-mentioned studies, along with the used (acoustic) models and the main findings.

Publication	Model	Result
Garnerin et al. (2019)	DNN-HMM	Male better
Garnerin et al. (2021)	End-to-end	Overall male better, increasing female training data reduces bias but does not reduce overall WER
Tatman (2017)	Youtube	Male better
Tatman and Kasten (2017)	Bing Youtube	No significant differences No significant differences
Sawalha and Abushariah (2013)	HMM	Female better
Adda-Decker and Lamel (2005)	HMM	Female better

Table 2: Overview of publications on gender bias

2.4.3 Age Bias

Intelligibility also differs across age groups due to variability in vocal characteristics. For example, the speech rate of elderly adults aged 65 and older is generally lower than that of younger adults aged 20–30 due to the "functional decline of elderly adults' vocal organs" (Kwon, Kim, & Choeh, 2016: p. 120). Additionally, elderly women were found to have longer silences between syllables and elderly men were found to produce significantly less energy in their second and third formants. However, the need or desire for voice assistants also exists in seniors, perhaps particularly in seniors. This is exemplified by Amazon's recent competition in which developers could win prize money for creating the best Alexa skill, designed specifically for people aged 55 and over (Hal Schwartz, 2022).

The recognition of speech by older adults has previously been researched in a broad range of settings. For instance, in 2008, Vipperla et al. published a longitudinal study in which ASR performance was compared across speakers' ages using recordings of the Supreme Court of the United States. They found that, in a context-dependent triphone HMM, the WER for speech by elderly speakers aged 59 through 87 was higher than that for adults under 55 years of age. In fact, a relative increase in WER of 31.3% was found. In addition, the model performed worse for female speakers in both age groups, although this is attributed to a lack of female speech in the training and test sets. When using speaker adaptation based on maximum likelihood linear regression, the overall WER for elderly speakers is reduced from 47.8% to 41.8%. Be that as it may, the discrepancy between the age groups is still present.

Comparable findings were reported in a subsequent study in which vocal tract length normalization (VTLN)—"a technique to transform the acoustic space of different speakers to a target speaker space" (Sivaraman, Mitra, Nam, Tiede, & Espy-Wilson, 2016: p. 455)—was used to in an effort to reduce interspeaker variability (Vipperla et al., 2010). Despite the improvement in overall performance, still a WER of 40.4% for elderly speakers and a 10

Similarly, a relative increase of 41% in WER was found in European Portuguese speech by speakers between the ages of 81 and 86 compared to 60-65-year-old speakers in an ASR system that combines HMMs with multi-layer perceptrons (Pellegrini et al., 2012). With acoustic models adapted to the age groups, however, the WER decreased, reducing the discrepancy between the younger and older age groups.

Many other approaches to elderly speech recognition have been adopted, for example by Hämäläinen et al. (2015). The AALFred Personal Life Assistant was trained for four languages:

French, Polish, Portuguese, and Hungarian. Depending on the language, Hämäläinen et al. attempted different models, though each acoustic model, originating from the Microsoft Speech Platform - Runtime, was optimized for elderly speech. The first model they included was a gender-dependent GMM trained with speech by younger people. Secondly, they implemented a deep belief network (DBN), which is a probabilistic generative model with multiple layers of stochastic hidden units (Hinton, 2009: para. 1). Lastly, a gender-independent GMM with speaker normalization was integrated. They found relative improvements in WER between 10% and 33%, with the best relative performance coming from the French DBN.

Furthermore, a range of hybrid DNN-HMM models was tested by Fukuda, Nishimura, Nishizaki, Iribe, and Kitaoka (2019) to investigate region-dependent ASR on a corpus of elderly Japanese speech of read-aloud news article sentences. The speakers were, on average, 79 years old and came from four different regions in Japan. Fukuda et al. compared the performance of models trained specifically on this corpus to the performance of models trained on speech by younger seniors with an average age of 68 years old and even younger adults aged 20–49. It was found that the best model performance was obtained by training the model on the speech of younger seniors. However, after adaptation to the region of the speakers, the models trained on the Corpus of Spontaneous Japanese (Furui, Maekawa, & Isahara, 2000) obtained lower WERs.

However, not only elderly speech deviates from the standard training datasets. Sawalha and Abushariah found that speakers over 30 years of age were recognized more poorly than those under 30. They ascribed this to younger speakers having "better vocal characteristics than the older speakers" (2013: para. 10).

Nonetheless, speech by children is highly variable depending on their age, showing higher fundamental frequencies (Busby & Plant, 1995) and a high magnitude for temporal and spectral acoustic parameters (Lee, Potamianos, & Narayanan, 1999). This variability and "obvious physiological differences" (Russell & D'Arcy, 2007: p. 108) cause children's voices to differ significantly from adults', making ASR more challenging for them, as is stated in the review of previous research into ASR for children's speech provided by Gerosa, Giuliani, Narayanan, and Potamianos (2009). Since then, this research has been further expanded. For example, Cosi (2009) transferred the SONIC speech recognizer (Pellom & Hacioglu, 2001) from English to Italian and was able to produce a phonetic error rate as low as 12.2% using techniques such as VTLN.

Similar results were found when Kathania et al. (2020) compared an ASR system with a GMM-HMM to a DNN-HMM to investigate the effect of formant modification. Both models showed an improvement over their respective baselines, from approximately 34% to 21% and 20% to 14% WER. They further inspected results by dividing the children test data into age groups, which showed a relative improvement of 45% for children aged 7 to 9. The model performed best for 10- to 14-year-olds. Though, in order to further reduce the overall WER, Kathania et al. combined their proposed method with VTLN and speaker rate adaptation and presented a WER of 12.35%.

The adaptation of a DNN from adult to child speech has also been demonstrated to significantly improve child speech recognition. Making use of Kullback-Leibler divergence regularization, weighted at the utterance level, Matassoni, Falavigna, and Giuliani (2016) were able to train a hybrid DNN-HMM model in Kaldi (Povey et al., 2011) using an adult speech corpus to recognize child speech with a WER as low as 10.6%.

The previously mentioned studies are outlined in Table 3, together with the evaluated (acoustic) models and the main findings of each study.

Publication	**Model**	**Result**
Vipperla et al. (2008)	HMM	Improvement after regression but still seniors worse
Vipperla et al. (2010)	HMM	Improvement after VTLN but still seniors worse
Pellegrini et al. (2012)	HMM	Improvement after age adaptation but seniors still worse
Hämäläinen et al. (2015)	Gender-dependent GMM	Improvement over the baseline
	DBN	Biggest improvement over the baseline
	Gender-independent GMM	Improvement over the baseline
Sawalha and Abushariah (2013)	HMM	Speakers over 30 years old worse
Cosi (2009)	HMM	Improvement over baseline after VTLN and adaptation
Kathania et al. (2020)	GMM-HMM	Improvement over the baseline
	DNN-HMM	Bigger improvement, biggest after VTLN and adaptation
Matassoni et al. (2016)	DNN-HMM	Improvement over baseline

Table 3: Overview of publications on age bias

2.4.4 Non-Nativeness Bias

As far as non-nativeness is concerned, Van Wijngaarden (2001) compared the intelligibility of both native and non-native speakers of Dutch, and found negative effects of non-nativeness on speech intelligibility. Especially Dutch vowels that do not occur in the non-native speakers' first language, American English, were recognized more poorly. Analogous to human listeners experiencing difficulty in understanding speech from non-native speakers (Van Wijngaarden, 2001), ASR systems show poorer performance for such speech. Therefore, researchers have been investigating various ways to improve the recognition of non-native speakers.

This is exemplified in a study by Aalburg and Hoege (2003), who simulated a non-native Spanish speech evaluation by training HMM-based models using peninsular Spanish as non-native and Colombian Spanish as native speech. They were able to improve model performance by adapting the acoustic models through probability density function-based clustering. Word recognition rates of approximately 80% and 96% were found for Spanish from Spain and Colombia, respectively. In other words, despite the overall improvements, the model performed more poorly for the simulated non-native Spanish (i.e., peninsular Spanish). In 2004, Aalburg and Hoege adopted a similar approach to improve the recognition of German as spoken by Turkish speakers, which showed a word recognition rate of 87.5% and 94.2% for non-native and native speakers, respectively.

In a recent study, Gretter, Matassoni, Falavigna, Evanini, and Leong (2020) presented an overview of a shared task on ASR for non-native child speech, specifically Italian learners of English. They showed that a model by Knill, Wang, Wang, Wu, and Gales (2020) using Kaldi (Povey et al., 2011) and HTK (Young et al., 2000) was able to reduce the baseline WER of 35% to just below 16%. It made use of graphemic TDNNs as the acoustic model. A 4-gram LM, standard unidirectional recurrent neural network language model (RNNLM) (Mikolov, Karafiát, Burget, Cernockỳ, & Khudanpur, 2010), as well as a four succeeding-word RNNLM (Chen, Liu, Qian, Gales, & Woodland, 2016; Chen, Liu, Wang, Gales, & Woodland, 2016) were used for rescoring.

In the following year, J. Wang, Zhu, Fan, Chu, and Alwan (2021) described their hybrid DNN-HMM model for Italian learners of German for the 2021 shared task, which included a BLSTM and RNNLM. They presented a non-speech state discriminative loss approach to avoid overfitting and found a WER of just under 40%, a relative improvement of around 12% compared to the baseline of 45%. However, the highest performance scores for this task were found using wav2vec (Baevski, Zhou, Mohamed, & Auli, 2020) and a transformer model for German, with a WER of 23.5%, and convolutional neural network TDNN for English, with a WER of 25.69% (Gretter et al., 2021).

A further attempt to improve the recognition of non-native speakers was carried out by Matassoni, Gretter, Falavigna, and Giuliani (2018). They applied transfer learning, i.e., generalizing a model that was developed for one task to another task (D. Wang & Zheng, 2015), in order to adapt a multilingual DNN to data from non-native speakers. The model was trained on Italian, German, and English child speech. It could be observed that the multilingual acoustic model was able to "compensate for the pronunciation differences of a non-native speaker" (p. 6232).

The above-mentioned studies are summarized in Table 4, which provides the evaluated (acoustic) models and main findings. These studies have thus attempted to improve the recognition of non-native speakers, but a discrepancy between non-native and native speakers still exists.

Publication	**Model**	**Result**
Aalburg and Hoege (2003)	HMM	Overall improvement but still simulated non-native worse
Aalburg and Hoege (2004)	HMM	Overall improvement but still non-native worse
Knill et al. (2020)	TDNN	Improvement over baseline
J. Wang et al. (2021)	BLSTM	Improvement over baseline
Gretter et al. (2021)	Wav2vec	Improvement over baseline
	Transformer	Improvement over baseline
Matassoni et al. (2018)	Multilingual DNN	Improvement over monolingual model

Table 4: Overview of publications on non-nativeness bias

3 Method

In this section I present the methods used in this research. Section 3.1 describes the two corpora needed for training and testing the ASR model. The model itself is defined in Section 3.2. In Section 3.3 I detail the process of the in-depth phoneme error analysis. The code used for the experiments in this thesis was based on the code used by Feng et al. (2021)[3] and is available at `https://github.com/Aariciah/jasmin`.

3.1 Corpora

3.1.1 Training

In order for this research to function as an extension of the study by Feng et al. (2021), the model of the current study is also trained using approximately 500 hours of speech from the Spoken Dutch Corpus (Oostdijk, 2000) (CGN, *Corpus Gesproken Nederlands*). The corpus contains

[3] The orginal code is available at `https://github.com/laurensw75/kaldi_egs_CGN` and `https://github.com/syfengcuhk/jasmin`

approximately 900 hours of speech, totaling just under nine million words, from various scenarios including spontaneous face-to-face conversations, news reports, and telephone conversations. The detailed component list (Van Eerten, 2007) is as follows:

- Spontaneous face-to-face conversations
- Interviews with teachers of Dutch
- Telephone dialogues recorded using a switchboard
- Telephone dialogues recorded using a mini disc recorder
- Simulated business negotiations
- Interviews and discussions broadcast on radio and television
- Discussions, debates, and meetings (especially in the political domain)
- Classes in secondary school (focusing on the teacher)
- Spontaneous commentaries, including sports, broadcast on radio and television
- Current affairs programs and news reports broadcast on radio and television
- News broadcasts on radio and television
- Commentaries broadcast on radio and television
- Sermons, lectures, and speeches
- Seminars, lectures, presentations
- Read-aloud texts

In addition, phonetic, syntactic, and prosodic annotations are available for some of the data, and speaker and audio file information can be found in the metadata. All speakers are over 18 years of age, with the eldest group being over 55 years old.

Although the CGN consists of speech from different regions in both the Netherlands and Belgium, only Netherlandic Dutch data is used to train the model. This allows for a clear view of the recognition of Flemish speech within a Dutch model.

3.1.2 Testing

To evaluate the model, Flemish data from the Jasmin-CGN corpus (Cucchiarini et al., 2006) was used. JASMIN stands for Jongeren, Anderstaligen en Machine Interactie voor het Nederlands (*Youngsters, Non-Dutch Speakers, and Machine Interaction for Dutch*); the corpus is an extension of the CGN that contains speech data of children and youngsters, older adults as well as speakers whose first language is not Dutch. Non-native speakers of Dutch in the Netherlands had a variety of first languages including Moroccan Arabic and Turkish. In Flanders, the non-native speakers were almost exclusively Francophones. Due to Belgium's multilingual landscape (i.e., Dutch in the North, French in the South, German in the East), many Belgians speak French at home. Two exceptions were included: one speaker whose first home language was Persian, and one speaker whose first home language was Dutch and whose second home language was Arabic. Furthermore, the corpus does not only feature read speech, but also includes speech from human-machine interactions. The inclusion of this mode in the corpus is important, seeing as many applications of ASR, such as smart kitchen devices (Kendrick, Frohnmaier, & Georges, 2021), are designed for (semi-) spontaneous speech rather than read speech. In addition, spontaneous or conversational speech is found to differ from read speech, for instance because of acoustic reduction (Schuppler, Ernestus, Scharenborg, & Boves, 2011) or just general variation in pronunciation (Schuppler, Adda-Decker, & Morales-Cordovilla, 2014).

Next to word-for-word transcriptions, automatic phonetic and syntactic annotations were added to the speech. The transcriptions were added automatically for native speakers and corrected manually for non-native speakers. The goal for the JASMIN corpus was to record approximately 95 hours of speech, of which two thirds in the Netherlands and one third in Flanders. Speech data from a total of 207 Flemish speakers was gathered.

The test data was into four regional categories according to the corpus metadata, which is based on the CGN metadata:

- West Flemish (W, peripheral region): 40 speakers, of which 22 female and 18 male
- East Flemish (E, transitional region): 38 speakers, of which 19 female and 19 male
- Brabantian (B, core region): 62 speakers, of which 32 female and 30 male
- Limburgish (L, peripheral region): 34 speakers, of which 18 female and 16 male

This was based on the area the participants lived in between the ages of 3 and 18. In the case of children, it was based on where they lived at the time of recording. For 33 speakers (of which 20 are female and 13 are male), no region label was provided.

In accordance with Feng et al. (2021) and the corpus aims, the age groups in this thesis are categorized as follows:

- Children aged 7–11: 43 speakers, of which 23 female and 20 male
- Youngsters aged 12–16: 44 speakers, of which 22 female and 22 male
- Older adults aged 60+: 38 speakers, of which 22 female and 16 male

More specifically, the ages of the older adults range from 65 to 89. Non-native speakers are divided into two age groups:

- Minors aged 7–16: 52 speakers, of which 25 female and 27 male
- Adults aged 18–60: 30 speakers, of which 19 female and 11 male

No 17-year-old speakers were included in the recordings.

For adult non-native speakers, information about their Dutch language proficiency was included in the form of their level in the Common European Framework of Reference for Languages (CEFR). The included levels are:

- A1 (Breakthrough): 9 speakers, of which 6 female and 3 male

- A2 (Waystage): 9 speakers, of which 5 female and 4 male
- B1 (Threshold): 11 speakers, of which 8 female and 3 male

The corpus metadata for the amount of time that non-native speakers have been learning Dutch ranges from four months to three years, although this information is not provided for all speakers. In the Jasmin-CGN corpus, the data is labeled with binary sexes: female and male.

- NC: native children between the ages of 7 and 11 → 6h 10m
- NY: native youngsters between the ages of 12 and 16 → 6h 10m
- NNC: non-native children and youngsters between the ages of 7 and 16 → 6h 10m
- NNA: non-native adults between the ages of 18 and 60 → 6h 10m
- NOA: native older adults ages 65 and over → 5h 5m

While nationality, education level, place of education, the number of years in the region, and the language spoken at home (including 'dialect'; 'tussentaal', which is "an intermediate variety between dialect and standard Dutch" (De Caluwe, 2009); and languages other than Dutch) are also present in the metadata, these are excluded from the analysis, as they are beyond the scope of this thesis.

3.2 ASR Model

The study makes use of a hybrid DNN-HMM system (Dahl et al., 2012) by Feng et al. (2021) using Kaldi (Povey et al., 2011). I adopt the TDNN-BLSTM model parameters from Feng et al. (2021). Three 1024-dimensional TDNN layers are complemented by three sets of bi-directional, 1024-dimensional LSTM layers. The lattice-free maximum mutual information criterion (Povey et al., 2016), which "performs sequence discrimination by introducing competing hypotheses through a denominator graph in the cost function" (Madikeri et al., 2020), is used to train the model, alongside various data augmentation methods to increase the amount of training data: noise (Snyder, Chen, & Povey, 2015), speed perturbation (Ko, Peddinti, Povey, & Khudanpur, 2015), and reverberation (Ko, Peddinti, Povey, Seltzer, & Khudanpur, 2017). High-resolution mel-frequency cepstral coefficients of 40 dimensions are used as input features to the acoustic model, which is trained for 4 epochs. A pre-trained GMM-HMM elicits context-dependent phone alignments through forced alignment, which are used to train the acoustic model. The system makes use of an RNNLM (Xu et al., 2018) with three TDNN and two LSTM layers. N-best results are generated by a tri-gram language model and rescored by the RNNLM, which are both trained on the transcriptions from the CGN.

3.3 Error Analysis

After training the acoustic model and rescoring the n-best results using an RNNLM, the models were evaluated by providing them with samples of both read and HMI speech by various groups as input. These groups are described in Section 3.1.2. The WERs were calculated as a whole as well as for each group separately, based on weighted averages. In other words, the amount of speech data per group impacted the final WER. For instance, fewer hours of speech are available for native older adults, so the WERs for their speech had a smaller influence on the total WER. The averages per speaker region were calculated separately by averaging the results of the different age groups, as there were no overarching regional groups created; therefore they are not weighted averages. Biases were then estimated by investigating the differences in WERs across the various speaker groups. A comparison was then made between the biases in the acoustic model and those in the rescored model.

Following that, the corpus transcriptions and the text given by the ASR model (i.e. the reference and recognition hypothesis, respectively) were transcribed phonetically, mapping the graphemes to phonemes, using the Phonemizer (Bernard & Titeux, 2021) with an eSpeak NG[4] backend. Subsequently, the transcribed files were aligned and PERs were calculated for each speaker group using the SCLITE Scoring Package from SCTK, the NIST Scoring Toolkit[5]. Manual calculations were made to combine the PERs for read speech and HMI. Likewise, the results of the different age groups per region were combined manually to obtain overarching regional results. For each speaker group, the top 5 misrecognized phonemes were compared.

4 Results

In this section, I present the detailed results of the ASR model performance across different groups of speakers of Flemish: first I report the results of the model before the RNNLM rescoring, then I report the results after the rescoring. Next, I describe the results of the in-depth phoneme error analysis.

4.1 Acoustic Model Results

The following tables provide a brief overview of WERs of the TDNN-BLSTM acoustic model without rescoring.

4.1.1 Regional and Age Bias

Differences across regional language variety can be seen in Tables 5 and 6 for read speech and HMI, respectively. Both tables demonstrate that West Flemish speakers were poorly recognized, with WERs as high as 60.44% for child HMI speech. Table 5 shows that a bias of 7.2% was found when comparing the average WERs for West Flemish and Brabantian. Similarly, Limburgish speakers were misrecognized, on average, 5.9% more than Brabantian speakers. Furthermore, in Table 6, a difference of nearly 20% can be observed between Brabantian children (52.41%) and Limburgish children (71.72%). The biases based on the average, when comparing with Brabant, were as follows: 6.65% against West Flanders, and 6.60% against Limburg. Overall, in both speech modes, the model performed best for Brabantian speech and worst for West Flemish speech.

However, the discrepancy between the regions decreased with the speakers' ages. The lowest WERs were found in older adults; in read speech, Brabantian seniors were misrecognized approximately 30% of the time, and in HMI East Flemish seniors were misrecognized 40% of the time. Seniors were recognized 17.27% better than children in read speech and 15.56% better in HMI.

[4] `https://github.com/espeak-ng/espeak-ng`

[5] `https://github.com/usnistgov/SCTK`

Group	W	E	B	L	W. Avg
NC	48.02	45.06	44.47	53.80	48.24
NY	48.02	37.40	38.32	44.76	42.66
NOA	35.24	29.77	27.25	29.99	30.97
Avg	43.88	37.41	36.68	42.85	

Table 5: Acoustic model WERs (%) for read speech across native speaker age groups and regions. "W.Avg" indicates the weighted average, and "Avg" indicates the regular average.

Group	W	E	B	L	W. Avg
NC	60.44	59.81	52.41	71.72	59.98
NY	53.31	43.63	44.52	43.97	47.22
NOA	46.38	41.83	43.44	44.38	44.42
Avg	53.41	48.42	46.79	53.36	

Table 6: Acoustic model WERs (%) for HMI speech across native speaker age groups and regions.

4.1.2 Gender Bias

Table 7 shows that overall, read speech is recognized better than HMI. A difference in performance across gender can also be observed; the model shows a lower WER for male speech in HMI compared to female speech, but the reverse is true for read speech. A 2.39% bias against male speech was found in read speech and a 0.77% bias against female speech was found for HMI.

	Read	HMI
F	45.10	52.23
M	47.49	51.46
W.Avg	46.20	51.90

Table 7: Acoustic model WERs (%) for read and HMI speech across genders. "F/M" indicates female and male.

4.1.3 Non-Nativeness Bias

Overall higher WERs can be seen in Table 8, which shows the model performance for non-native speakers. Here too, the model performed better for adult speech than for child speech. However, in all four cases, the average WERs are higher than for native speakers of the respective categories. A difference in gender can also be observed; in the read mode female speech was recognized more poorly, whereas in HMI, girls were better recognized than boys, but men were better recognized than women.

A comparison of read speech by non-native children (aged 7–16) with read speech by native children (aged 7–11) showed that there was a bias of 10.95% against non-native children.

Group	Read			HMI		
	F	**M**	**W. Avg**	**F**	**M**	**W. Avg**
NNC	60.11	58.75	59.19	60.07	60.98	61.00
NNA	52.24	47.57	50.24	58.30	52.23	56.06

Table 8: Acoustic model WERs (%) for read and HMI speech across non-native speakers.

4.2 Results with RRNLM Rescoring

4.2.1 Regional and Age Bias

To ensure a more accurate and complete view of the ASR model performance, the following tables display the results after passing through the RNNLM for rescoring. In all instances, the WERs decreased compared to the results without RNNLM rescoring.

Tables 9 and 10 show the WERs for each region and the three native speaker age groups. Additionally, they show the WERs per gender (female, male). The final column in both tables contains the weighted average per age group, whereas the bottom row provides the regular average per region.

Table 9 demonstrates that, for read speech, speakers from Brabant were best recognized overall, with the lowest WER being found for older adults. However, for native youngsters, the best recognition was found in East Flemish speakers. Nonetheless, the difference between the WERs for East Flemish youngsters and Brabantian youngsters was merely 0.1%. The highest WER was found for Limburgish child speakers (50.25%), with an even higher rate for boys (62.52%). In almost all cases, male speakers were recognized more poorly than female speakers.

A quantification of the bias shows that the model performed 7.08% worse for West Flemish and 6.61% for Limburgish speakers compared to Brabantian speakers. In addition, a bias of 17.23% was found against children, compared to seniors.

From Table 10 it is clear that speech from HMI was recognized more poorly than read speech, with WERs approximately 10% higher than in Table 9. This discrepancy can be found across the different regions. Still, overall, the model performed best for Brabantian speech (42.15%) compared to the other regions. However, older adults from East Flanders and youngsters from Limburg were recognized better than their Brabantian peers. When comparing the averages per region, a 6.55% bias against West Flemish speakers and 6.03% against Limburgish speakers was found.

Like Table 9, Table 10 shows that the model performed best on speech by older adults, with a weighted average of 40.42%. This resulted in a bias against children of 15.12%. In most cases, male speech was recognized more poorly, though this trend is not as clear as in Table 9.

4.2.2 Gender Bias

While the previous tables provide detailed results for male and female speakers across different regions and age groups, Table 11 shows the overall WERs per gender and mode of speech. Compared to the model without RNNLM rescoring, the average WERs for read speech and HMI speech decreased from 46.2% and 51.9% to 41.97% and 47.90% respectively. Contrary to Table 7, female speech was better recognized than male speech in both modes, resulting in a bias against male speakers of 2.71% in read speech and 0.33% in HMI.

Group	**W**	**E**	**B**	**L**	**W. Avg**
NC	43.75	41.14	40.06	50.25	44.06
	42.40 , 45.35	39.62 , 43.45	39.84 , 40.30	39.06, 62.52	40.22 , 48.71
NY	41.42	31.58	31.68	39.07	36.68
	40.30 , 42.81	26.14 , 38.90	33.79 , 29.36	33.33 , 44.30	34.65 , 38.99
NOA	31.03	25.53	23.22	25.45	26.83
	29.04 , 34.36	25.21 , 26.14	22.82 , 23.83	29.18 , 21.41	26.64 , 27.02
Avg	38.73	32.74	31.65	38.26	
	37.25 , 40.84	30.32 , 36.16	32.15 , 31.16	33.86 , 42.74	

Table 9: Rescored model WERs (%) for read speech across native speaker age groups and regions. Weighted averages per group are at the top. The bottom results are in the format: female, male

Group	**W**	**E**	**B**	**L**	**W. Avg**
NC	55.67	54.62	48.60	67.69	55.54
	54.71 , 56.78	53.96 , 55.77	46.88 , 49.68	61.11 , 72.31	53.77 , 57.10
NY	47.54	40.57	39.57	36.85	42.13
	47.14 , 48.26	38.35 , 45.15	39.16 , 40.00	33.88 , 40.30	40.96 , 44.09
NOA	42.89	38.16	38.28	40.00	40.42
	38.42 , 48.96	46.35 , 30.79	36.67 , 40.86	43.36 , 36.39	40.48 , 39.79
Avg	48.70	44.45	42.15	48.18	
	46.76 , 51.33	46.22 , 43.90	40.90 , 43.51	46.12 , 49.67	

Table 10: Rescored model WERs (%) for HMI speech across native speaker age groups and regions. Weighted average per group are at the top. The bottom results are in the format: female, male.

4.2.3 Non-Nativeness Bias

Table 12 shows comparable results as Table 8; read speech by male non-native speakers was better recognized than that of their female counterparts, but the model performed better for non-native girls in HMI. It can also be observed that, for the most part, the same age bias is present as in the results for native speakers: children were more often misrecognized than adults. While the reverse is true for girls and women in HMI, the difference in WERs amounts to merely 0.03%.

Non-native children's read speech (aged 7–16) were recognized 11.89% worse than native children's (aged 7–11).

For non-native speakers, the difference in performance was the largest in speakers with the lowest proficiency level (A1). However, contrary to expectations, the performance did not steadily improve for the levels A2 and B1. While the WERs for A2 speakers were lower than

	Read	**HMI**
F	40.83	47.69
M	43.54	48.02
W. Avg	41.97	47.90

Table 11: Rescored model WERs (%) for read and HMI speech across genders.

Group	**Read**			**HMI**		
	F	**M**	**W. Avg**	**F**	**M**	**W. Avg**
NNC	57.23	32.75	55.95	54.45	56.87	56.68
NNA	49.26	44.45	47.28	54.48	49.67	52.61

Table 12: Rescored model WERs (%) for read and HMI speech across non-native speakers.

those for A1 speakers, a higher WER was found for speakers in the B1 level. Table 13 thus shows that the best performance for non-native speakers was obtained for speakers in the A2 level. This is true for both read and HMI speech. The overall lowest WERs for non-native adults were found for female speakers at the A2 level.

This does not mean, however, that female non-native speech was recognized better overall. In fact, for read speech, the model showed poorer performance for the A1 and B1 levels. Nonetheless, a larger discrepancy between female and male non-native speakers can be found in HMI speech at the A1 level, where the WER for women lies at approximately 60% and the WER for men at approximately 74%. The weighted averages in Table 10 were skewed toward the scores for female speech, as this part of the corpus contains more female speakers. Be that as it may, Table 13 shows that, in general, approximately one in two words was not recognized correctly by the ASR model.

CEFR	**Read**			**HMI**		
	F	**M**	**W. Avg**	**F**	**M**	**W. Avg**
A1	52.69	50.17	51.79	59.95	74.07	62.09
A2	40.13	**46.43**	**42.88**	42.84	46.51	43.78
B1	51.97	46.35	50.27	59.76	50.25	55.02

Table 13: Rescored model WERs (%) for read and HMI speech across language learning levels.

4.3 Phoneme Error Analysis

Looking at the data in general, i.e., the data of all groups and both read and HMI speech, the following five phonemes had the highest PERs: /eː/, /ɣ/, /ə/, /ɛ/. Most of these are also among the worst recognized phonemes in the different speaker groups. Across all regions, /eː/ is the phoneme with the most misrecognitions. For West Flemish speakers the following phonemes were most often misrecognized: /eː/, /ɛ/, /ə/, /t/, /ɑ/. In East Flemish speech the following phonemes had the highest PERs: /eː/, /ə/, /ɛ/, /h/, /ɣ/. Brabantian speakers were recognized poorly when pronouncing /eː/, /ə/, /ɛ/, /t/, and /n/. For the Limburg region, the following phonemes were found to have the highest PERs: /eː/, /ɣ/, /ə/, /t/, /ɛ/.

In female speech, /eː/, /ɛ/, /ɣ/, /ə/, and /t/ were poorly recognized. The phonemes with the highest PER in male speech were the same as the overall worst recognized phonemes. The same phonemes were also poorly recognized when spoken by Flemish children, although the order is slightly different: /eː/, /ɣ/, /h/, /ɛ/, /ə/. The model performed poorly for /ə/, /eː/, /ɛ/, /t/, and /n/ as spoken by Flemish youngsters. The pronunciation of /eː/, /t/, /ɛ/, /ə/, and /ɑ/ by Flemish seniors was also most often misrecognized.

For non-native children, the following phonemes were found to be most poorly recognized: /eː/, /h/, /ɣ/, /ə/, /ɛ/ For non-native adults, however, /eː/, /ɣ/, /ə/, /ɛ/, and /n/ had the highest PERs. These results demonstrate that there was no clear difference in the recognition of certain phonemes across regions, genders, age groups, or native versus non-native speakers.

5 Discussion and Conclusion

This study's main aim was to examine whether a Dutch hybrid DNN-HMM ASR model would show performance biases, i.e., differences in WER, when comparing speakers from different regions, of different genders, in different age groups, and against non-native speakers of Dutch, using data from Flanders. Furthermore, it sought to determine whether the biases decreased after applying RNNLM rescoring. The speech data used to train the models came from the Spoken Dutch Corpus (Oostdijk, 2000), though only the Netherlandic Dutch section was used. In order to evaluate the model, Flemish data from the JASMIN-CGN corpus (Cucchiarini et al., 2006) was used and divided into different speaker groups. WERs were compared across these speaker groups and phoneme errors were reviewed.

The study expanded on previous research by Feng et al. (2021), who found that the model performed worse for Flemish speech than for regions in the Netherlands. Upon closer analysis of the Flemish regions, interesting results were found. In line with studies into human intelligibility that have shown that West Flemish and Limburgish speech were perceived as less intelligible (Impe & Geeraerts, 2008; Van Bezooijen & Van den Berg, 1999a; 1999b), the model performance was poorest for West Flemish speakers, followed by Limburgish speakers. The overall biases for these regions were approximately 6.5%. Both regions are considered peripheral regions, whereas Brabant (including Antwerp) is considered a central region and obtained the best performance.

However, the performance did vary based on the speakers' ages. While it was expected that native youngsters would have the lowest WERs, as found in the Netherlandic Dutch data (Feng et al., 2021), on average, seniors were recognized better. The obtained WERs for older adults were relatively similar to those found by Feng et al. (2021), but the results for youngsters and children were considerably higher. In fact, a difference in WER of approximately 17% was found when comparing native older adults to native children. These results are quite surprising considering the various attempts to improve child speech recognition that have previously been undertaken (e.g., Kathania et al., 2020; Matassoni et al., 2016).

Additionally, Feng et al. (2021) found that female speech was recognized better than male speech. Their results align with the findings from this experiment; overall male speakers were misrecognized more than female speakers. Nonetheless, the bias against male speakers only amounts to 2.7%, indicating a better balance in recognition than for the different age groups. Gender bias seems to be highly susceptible to the training data, as previous studies have found varying results. For instance Garnerin et al. (2019) found that an imbalanced corpus performed better for male speakers, whereas Adda-Decker and Lamel (2005) obtained better recognition for female speakers in a balanced corpus.

In line with the findings from Feng et al. (2021), poor recognition results were obtained for non-native speakers. When comparing the results for non-native children to those of native children, a bias of approximately 11% was found. Taking a closer look at the CEFR levels of adult non-native speakers (A1, A2, B1) showed that, contrary to expectations, the model did not perform the best for speakers in the B1 level. Instead, the WERs for speakers in the A2 level were approximately 10% lower. One possible explanation may be that measuring language learning levels depends on a number of factors, usually speaking, writing, and reading. Therefore, it is possible that the overall level of the speaker is assessed at B1, whereas the speaking skills are assessed at A2, for instance.

A quantification of the biases in the model before and after RNNLM rescoring showed that along with an overall decrease in WERs, the difference in WERs between the Flemish regions became smaller. However, unexpectedly, this was not the case for the age biases, which stayed approximately the same. Moreover, the biases against male and non-native children slightly increased. These results warrant further investigations into biases in language models.

Finally, the examination of misrecognized phonemes uncovered that /eː/, /ɣ/, /ə/, /ɛ/, and /h/ were most often misrecognized across the whole test dataset. These remained relatively constant across the speaker groups. I did not replicate the misrecognitions for Flanders found in Feng et al. (2021), namely /y/, /œy/, and /ɑu/. This might be due to the use of another automatic phonetic transcription and phoneme error calculation.

In conclusion, in this thesis I presented an experiment on bias due to regional language varieties, age, gender, and non-nativeness in Flemish automatic speech recognition using a Dutch model. By investigating WERs and phoneme misrecognitions across different groups, I provided an overview of biases against various groups of Flemish speakers in Dutch ASR. While the findings provide invaluable insights into ASR model performance discrepancies between and within the various investigated groups, I am aware of the limited scope of the test data. While approximately six hours of speech samples per age group may generally be sufficient, dividing the groups according to gender and then further according to region may impact the generalizability of the results. Therefore, I believe that future research into bias in Flemish ASR, where a larger dataset is used to test the model, could yield more accurate and generalizable results. Furthermore, this study could be expanded on by researching biases in a DNN-HMM model trained on Flemish speech or a combination of Netherlandic Dutch and Flemish speech. Investigations into Flemish biases in other models, such as transformer models, are also recommended. Nonetheless, the findings have important implications for the development of inclusive ASR so that Flemish speakers can also make use of Dutch ASR, regardless of their region, gender, age, or native language. Moreover, the research contributes to a growing body of studies on ASR bias and thus provides a stronger foundation for future research.

6 References

Aalburg, S., & Hoege, H. (2003). Approaches to foreign-accented speaker-independent speech recognition. In *Proceedings of the 8th European Conference on Speech Communication and Technology (Eurospeech 2003)* (pp. 1489–1492).

Aalburg, S., & Hoege, H. (2004). Foreign-accented speaker-independent speech recognition. In *Proceedings of Interspeech 2004* (pp. 1465–1468). doi: 10.21437/Interspeech.2004-558

Abushariah, M. A.-A. M., Ainon, R. N., Zainuddin, R., Alqudah, A. A. M., Elshafei Ahmed, M., & Khalifa, O. O. (2012). Modern standard Arabic speech corpus for implementing and evaluating automatic continuous speech recognition systems. *Journal of the Franklin Institute*, *349*(7), 2215–2242. Retrieved from `https://www.sciencedirect.com/science/article/pii/S0016003211001013` doi: https://doi.org/10.1016/j.jfranklin.2011.04.011

Adda-Decker, M., & Lamel, L. (2005). Do speech recognizers prefer female speakers? In *Proceedings of Interspeech 2005* (pp. 2205–2208). doi: 10.21437/Interspeech.2005-699

Auxier, B. (2019). 5 things to know about Americans and their smart speakers. *Pew Research Center*. Retrieved 2022-04-23, from `https://www.pewresearch.org/fact-tank/2019/11/21/5-things-to-know-about-americans-and-their-smart-speakers/`

Bacchiani, M., Beaufays, F., Schalkwyk, J., Schuster, M., & Strope, B. (2008). Deploying goog-411: Early lessons in data, measurement, and testing. In *Proceedings of the 2008 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP 2008)* (pp. 5260–5263). doi: 10.1109/ICASSP.2008.4518846

Baevski, A., Zhou, Y., Mohamed, A., & Auli, M. (2020). wav2vec 2.0: A framework for self-supervised learning of speech representations. *Advances in Neural Information Processing Systems*, *33*, 12449–12460.

Baum, M., Erbach, G., Kommenda, M., Niklfeld, G., & Puig-Waldmüller, E. (2001). Speech and multimodal dialogue systems for telephony applications based on a speech database of Austrian German. *ÖGAI-Journal*, *20*, 29–34.

Baum, M., Muhr, R., & Kubin, G. (2001). A phonetic lexicon for adaptation in ASR for Austrian German. In *ISCA Tutorial and Research Workshop (ITRW) on Adaptation Methods for Speech Recognition.*

Benzeghiba, M., De Mori, R., Deroo, O., Dupont, S., Erbes, T., Jouvet, D., ... Wellekens, C. (2007, oct). Automatic speech recognition and speech variability: A review. *Speech Communication*, *49*(10–11), 763–786. Retrieved from `https://doi.org/10.1016/j.specom.2007.02.006` doi: 10.1016/j.specom.2007.02.006

Beringer, N., Schiel, F., & Regel-Brietzmann, P. (1998). German regional variants - a problem for automatic speech recognition? In *Proceedings of the 5th International Conference on Spoken Language Processing (ICSLP 1998)* (pp. 85–88).

Bernard, M., & Titeux, H. (2021). Phonemizer: Text to phones transcription for multiple languages in python. *Journal of Open Source Software*, *6*(68), 3958. Retrieved from `https://doi.org/10.21105/joss.03958` doi: 10.21105/joss.03958

Blodgett, S. L., Barocas, S., Daumé, H., & Wallach, H. (2020). *Language (technology) is power: A critical survey of "bias" in NLP.* arXiv. Retrieved from `https://arxiv.org/abs/2005.14050` doi: 10.48550/ARXIV.2005.14050

Busby, P. A., & Plant, G. L. (1995). Formant frequency values of vowels produced by preadolescent boys and girls. *The Journal of the Acoustical Society of America*, *97*(4), 2603–2606.

Chang, J., Arbeláez, P., Switz, N., Reber, C., Tapley, A., Davis, J. L., ... Malik, J. (2012). Automated tuberculosis diagnosis using fluorescence images from a mobile microscope. In N. Ayache, H. Delingette, P. Golland, & K. Mori (Eds.), *Medical image computing and computer-assisted intervention – miccai 2012* (pp. 345–352). Berlin, Heidelberg: Springer Berlin Heidelberg.

Chen, X., Liu, X., Qian, Y., Gales, M. J. F., & Woodland, P. C. (2016). CUED-RNNLM — an open-source toolkit for efficient training and evaluation of recurrent neural network language models. In *Proceedings of the 2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)* (pp. 6000–6004). doi: 10.1109/

ICASSP.2016.7472829

Chen, X., Liu, X., Wang, Y., Gales, M. J. F., & Woodland, P. C. (2016). Efficient training and evaluation of recurrent neural network language models for automatic speech recognition. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, *24*(11), 2146–2157. doi: 10.1109/TASLP.2016.2598304

Cheng, L. S. P., Burgess, D., Vernooij, N., Solís-Barroso, C., McDermott, A., & Namboodiripad, S. (2021). The problematic concept of native speaker in psycholinguistics: Replacing vague and harmful terminology with inclusive and accurate measures. *Frontiers in Psychology*, *12*. Retrieved from https://www.frontiersin.org/articles/10.3389/fpsyg.2021.715843 doi: 10.3389/fpsyg.2021.715843

Class, F., Kaltenmeier, A., & Regel-Brietzmann, P. (1993). Optimization of an hmm-based continuous speech recognizer. In *Proceedings of the 3rd European Conference on Speech Communication and Technology (Eurospeech 1993)* (pp. 803–806).

Contreras Carrasco, O. (2019). *Gaussian mixture models explained.* https://towardsdatascience.com/gaussian-mixture-models-explained-6986aaf5a95. (Accessed: 2022-08-02)

Cosi, P. (2009). On the development of matched and mismatched Italian children's speech recognition systems. In *Proceedings of Interspeech 2009* (pp. 540–543). doi: 10.21437/Interspeech.2009-195

Cucchiarini, C., Van hamme, H., van Herwijnen, O., & Smits, F. (2006, May). JASMIN-CGN: Extension of the Spoken Dutch Corpus with speech of elderly people, children and non-natives in the human-machine interaction modality. In *Proceedings of the Fifth International Conference on Language Resources and Evaluation (LREC'06).* Genoa, Italy: European Language Resources Association (ELRA). Retrieved from http://www.lrec-conf.org/proceedings/lrec2006/pdf/254_pdf.pdf

Dahl, G. E., Yu, D., Deng, L., & Acero, A. (2012). Context-dependent pre-trained deep neural networks for large-vocabulary speech recognition. *IEEE Transactions on audio, speech, and language processing*, *20*(1), 30–42.

Davis, K. H., Biddulph, R., & Balashek, S. (1952). Automatic recognition of spoken digits. *The Journal of the Acoustical Society of America*, *24*(6), 637-642. Retrieved from https://doi.org/10.1121/1.1906946 doi: 10.1121/1.1906946

De Caluwe, J. (2009). Tussentaal wordt omgangstaal in vlaanderen [Journal Article]. *Nederlandse Taalkunde*, *14*(1), 8–25. Retrieved from https://www.aup-online.com/content/journals/10.5117/NEDTAA2009.1.TUSS339 doi: https://doi.org/10.5117/NEDTAA2009.1.TUSS339

De Caluwe, J., Hüning, M., Vogl, U., & Moliner, O. (2012). Dutch in Belgium facing multilingualism. *Standard Languages and Multilingualism in European History*, *1*, 259–282.

Delarue, S., & De Caluwe, J. (2015). Eliminating social inequality by reinforcing standard language ideology? Language policy for Dutch in Flemish schools. *Current Issues in Language Planning*, *16*(1-2), 8–25. Retrieved from https://doi.org/10.1080/14664208.2014.947012 doi: 10.1080/14664208.2014.947012

Devos, M. (2006). Genese en structuur van het Vlaamse dialectlandschap. In *Structuren in talige variatie in Vlaanderen* (pp. 35–61). Academia Press.

Dizdarevic, V., Hagmuller, M., Kubin, G., Pernkopf, E., & Baum, M. (2004). Prosody-based recognition of spoken German varieties. In *Proceedings of the 2004 IEEE International Conference on Acoustics, Speech, and Signal Processing* (Vol. 1, pp. 929–932). doi: 10.1109/ICASSP.2004.1326139

Doukhan, D., Poels, G., Rezgui, Z., & Carrive, J. (2018, December). Describing gender equality in French audiovisual streams with a deep learning approach. *VIEW Journal*, *7*(14), 103–122. doi: 10.18146/2213-0969.2018.jethc156

Feine, J., Gnewuch, U., Morana, S., & Maedche, A. (2019). Gender bias in chatbot design. In *International Workshop on Chatbot Research and Design* (pp. 79–93).

Feng, S., Kudina, O., Halpern, B. M., & Scharenborg, O. (2021). Quantifying bias in automatic speech recognition. *arXiv preprint arXiv:2103.15122*.

Fukuda, M., Nishimura, R., Nishizaki, H., Iribe, Y., & Kitaoka, N. (2019). A new corpus of elderly Japanese speech for acoustic modeling, and a preliminary investigation of dialect-dependent speech recognition. In *2019 22nd conference of the oriental cocosda international committee for the co-ordination and standardisation of speech databases and assessment techniques (o-cocosda)* (pp. 1–6). doi: 10.1109/O-COCOSDA46868.2019.9041216

Furui, S. (2005). 50 years of progress in speech and speaker recognition. In *Proceedings of the 10th International Conference on Speech and Computer (SPECOM 2005)* (pp. 1–9).

Furui, S. (2010). History and development of speech recognition. In *Speech technology* (pp. 1–18). Springer.

Furui, S., Maekawa, K., & Isahara, H. (2000). A Japanese national project on spontaneous speech corpus and processing technology. In *ASR2000-Automatic Speech Recognition: Challenges for the new Millenium ISCA Tutorial and Research Workshop (ITRW).*

Garnerin, M., Rossato, S., & Besacier, L. (2019). 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* (pp. 3–9). New York, NY, USA: Association for Computing Machinery. Retrieved from https://doi.org/10.1145/3347449.3357480 doi: 10.1145/3347449.3357480

Garnerin, M., Rossato, S., & Besacier, L. (2020). Pratiques d'évaluation en ASR et biais de performance. In G. Adda, M. Amblard, & K. Fort (Eds.), *6e conférence conjointe Journées d'Études sur la Parole (JEP, 33e édition), Traitement Automatique des Langues Naturelles (TALN, 27e édition), Rencontre des Étudiants Chercheurs en Informatique pour le Traitement Automatique des Langues (RÉCITAL, 22e édition). 2e atelier Éthique et TRaitemeNt Automatique des Langues (ETeRNAL)* (pp. 1–9). Nancy, France: ATALA. Retrieved from https://hal.archives-ouvertes.fr/hal-02750220

Garnerin, M., Rossato, S., & Besacier, L. (2021, August). Investigating the impact of gender representation in ASR training data: a case study on Librispeech. In *3rd Workshop on Gender Bias in Natural Language Processing* (pp. 86–92). Online, France: Association for Computational Linguistics. Retrieved from https://hal.univ-grenoble-alpes.fr/hal-03472117 doi: 10.18653/v1/2021.gebnlp-1.10

Geeraerts, D., Grondelaers, S., & Speelman, D. (1999). *Convergentie en divergentie in de nederlandse woordenschat: een onderzoek naar kleding-en voetbaltermen.* PJ Meertens-Instituut; Amsterdam.

Gerosa, M., Giuliani, D., Narayanan, S., & Potamianos, A. (2009). A review of ASR technologies for children's speech. In *Proceedings of the 2nd Workshop on Child, Computer and Interaction* (pp. 1–8). New York, NY, USA: Association for Computing Machinery. Retrieved from https://doi.org/10.1145/1640377.1640384 doi: 10.1145/1640377.1640384

Gretter, R., Matassoni, M., Falavigna, D., Evanini, K., & Leong, C. W. (2020). Overview of the Interspeech TLT2020 Shared Task on ASR for Non-Native Children's Speech. In *Proceedings of Interspeech 2020* (pp. 245–249). doi: 10.21437/Interspeech.2020-2133

Gretter, R., Matassoni, M., Falavigna, D., Misra, A., Leong, C., Knill, K., & Wang, L. (2021). ETLT 2021: Shared Task on Automatic Speech Recognition for Non-Native Children's Speech. In *Proceedings of Interspeech 2021* (pp. 3845–3849). doi: 10.21437/Interspeech.2021-1237

Hal Schwartz, E. (2022). *Amazon opens $45,000 Alexa skills for seniors competition.* https://voicebot.ai/2022/07/22/amazon-opens-45000-alexa-skills-for-seniors-competition/. (Accessed: 2022-08-03)

Hämäläinen, A., Teixeira, A., Almeida, N., Meinedo, H., Fegyó, T., & Dias, M. S. (2015). Multilingual speech recognition for the elderly: The AALFred personal life assistant. *Procedia Computer Science*, *67*, 283–292.

Hamidi, M., Satori, H., Zealouk, O., & Satori, K. (2020). Amazigh digits through interactive speech recognition system in noisy environment. *International Journal of Speech Technology*, *23*(1), 101–109.

Hellström, T., Dignum, V., & Bensch, S. (2020). *Bias in machine learning – what is it good for?* arXiv. Retrieved from `https://arxiv.org/abs/2004.00686` doi: 10.48550/ARXIV.2004.00686

Hinton, G. E. (2009). Deep belief networks. *Scholarpedia*, *4*(5), 5947. (revision #91189) doi: 10.4249/scholarpedia.5947

Hochreiter, S., & Schmidhuber, J. (1997). Long short-term memory. *Neural Computation*, *9*(8), 1735-1780. doi: 10.1162/neco.1997.9.8.1735

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

IBM. (n.d.-a). *IBM Shoebox.* `https://www.ibm.com/ibm/history/exhibits/specialprod1/specialprod1_7.html`. (Accessed: 2022-08-01)

IBM. (n.d.-b). *Pioneering speech recognition.* `https://www.ibm.com/ibm/history/ibm100/us/en/icons/speechreco/`. (Accessed: 2022-08-01)

Impe, L., & Geeraerts, D. (2008). Babel in Vlaanderen - Een experimenteel onderzoek naar onderlinge verstaanbaarheid tussen Vlamingen. *Studies van de BKL*, *3*.

Impe, L., Geeraerts, D., & Speelman, D. (2008). Mutual intelligibility of standard and regional Dutch language varieties. *International Journal of Humanities and Arts Computing*, *2*(1-2), 101–117.

Ji, S., Han, X., Hou, Y., Song, Y., & Du, Q. (2020). Remaining useful life prediction of airplane engine based on pca–blstm. *Sensors*, *20*(16), 4537.

Kamath, U., Liu, J., & Whitaker, J. (2019). *Deep learning for nlp and speech recognition* (Vol. 84). Springer.

Kathania, H. K., Kadiri, S. R., Alku, P., & Kurimo, M. (2021). Using data augmentation and time-scale modification to improve asr of children's speech in noisy environments. *Applied Sciences*, *11*(18). Retrieved from `https://www.mdpi.com/2076-3417/11/18/8420` doi: 10.3390/app11188420

Kathania, H. K., Reddy Kadiri, S., Alku, P., & Kurimo, M. (2020). Study of formant modification for children ASR. In *Proceedings of the 2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)* (pp. 7429–7433). doi: 10.1109/ICASSP40776.2020.9053334

Kendrick, C., Frohnmaier, M., & Georges, M. (2021, 12–13 November). Audio-visual recipe guidance for smart kitchen devices. In *Proceedings of The Fourth International Conference on Natural Language and Speech Processing (ICNLSP 2021)* (pp. 257–261). Trento, Italy: Association for Computational Linguistics. Retrieved from `https://aclanthology.org/2021.icnlsp-1.30`

Kinsella, B. (2022). *Over half of U.S. adults have smart home devices, nearly 30% use voice assistants with them – new report.* `https://voicebot.ai/2022/06/20/over-half-of-u-s-adults-have-smart-home-devices-nearly-30-use-voice-assistants-with-them-new-report/`. (Accessed: 2022-08-01)

Knill, K. M., Wang, L., Wang, Y., Wu, X., & Gales, M. J. (2020). Non-native children's automatic speech recognition: The INTERSPEECH 2020 Shared Task ALTA Systems. In *Proceedings of Interspeech 2020* (pp. 255–259). doi: 10.21437/Interspeech.2020-2154

Ko, T., Peddinti, V., Povey, D., & Khudanpur, S. (2015). Audio augmentation for speech recognition. In *Proceedings of Interspeech 2015.*

Ko, T., Peddinti, V., Povey, D., Seltzer, M. L., & Khudanpur, S. (2017). A study on data augmentation of reverberant speech for robust speech recognition. In *Proceedings of the 2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)* (pp. 5220–5224).

Kumar, S. S. (2018). *Acoustic modeling (ASR part 2).* `hhttps://sujayskumar.com/2018/12/`

05/asr-part2/. (Accessed: 2022-08-02)

Kwon, S., Kim, S.-J., & Choeh, J. Y. (2016). Preprocessing for elderly speech recognition of smart devices. *Computer Speech & Language*, *36*, 110–121.

Lee, S., Potamianos, A., & Narayanan, S. (1999). Acoustics of children's speech: Developmental changes of temporal and spectral parameters. *The Journal of the Acoustical Society of America*, *105*(3), 1455–1468.

Lehr, M., Gorman, K., & Shafran, I. (2014). Discriminative pronunciation modeling for dialectal speech recognition. In *Proceedings of Interspeech 2014* (pp. 1458–1462).

Lowerre, B. T. (1976). *The HARPY speech recognition system.* Carnegie Mellon University.

Madikeri, S., Khonglah, B. K., Tong, S., Motlicek, P., Bourlard, H., & Povey, D. (2020). Lattice-free maximum mutual information training of multilingual speech recognition systems. In *Proceedings of Interspeech 2020* (pp. 4746–4750). doi: 10.21437/Interspeech.2020-2919

Matassoni, M., Falavigna, D., & Giuliani, D. (2016). DNN adaptation for recognition of children speech through automatic utterance selection. In *Proceedings of the 2016 IEEE Spoken Language Technology Workshop (SLT)* (pp. 644–651). doi: 10.1109/SLT.2016.7846331

Matassoni, M., Gretter, R., Falavigna, D., & Giuliani, D. (2018). Non-native children speech recognition through transfer learning. In *Proceedings of the 2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)* (pp. 6229–6233). doi: 10.1109/ICASSP.2018.8462059

Mehrabi, N., Morstatter, F., Saxena, N., Lerman, K., & Galstyan, A. (2021, jul). A survey on bias and fairness in machine learning. *ACM Computing Surveys (CSUR)*, *54*(6), 1–35. Retrieved from https://doi.org/10.1145/3457607 doi: 10.1145/3457607

Mikolov, T., Karafiát, M., Burget, L., Cernockỳ, J., & Khudanpur, S. (2010). Recurrent neural network based language model. In *Proceedings of Interspeech 2010* (Vol. 2, pp. 1045–1048).

Mitchell, T. M. (1997). *Machine learning* (Vol. 1) (No. 9). McGraw-Hill.

Nigmatulina, I., Kew, T., & Samardzic, T. (2020, December). ASR for non-standardised languages with dialectal variation: The case of Swiss German. In *Proceedings of the 7th workshop on NLP for similar languages, varieties and dialects* (pp. 15–24). Barcelona, Spain (Online): International Committee on Computational Linguistics (ICCL). Retrieved from https://aclanthology.org/2020.vardial-1.2

Odijk, J. (2012). *Het Nederlands in het Digitale Tijdperk – The Dutch Language in the Digital Age.* Springer. (Available online at http://www.meta-net.eu/whitepapers)

Oostdijk, N. (2000, May). The spoken Dutch corpus. Overview and first evaluation. In *Proceedings of the Second International Conference on Language Resources and Evaluation (LREC'00).* Athens, Greece: European Language Resources Association (ELRA). Retrieved from http://www.lrec-conf.org/proceedings/lrec2000/pdf/110.pdf

O'Shaughnessy, D. (2000). *Speech communications: Human and machine.* IEEE Press.

O'Shaughnessy, D. (2019). Recognition and processing of speech signals using neural networks. *Circuits, Systems, and Signal Processing*, *38*(8), 3454–3481.

Panayotov, V., Chen, G., Povey, D., & Khudanpur, S. (2015). Librispeech: An ASR corpus based on public domain audio books. In *Proceedings of the 2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)* (pp. 5206–5210). doi: 10.1109/ICASSP.2015.7178964

Pellegrini, T., Trancoso, I., Hämäläinen, A., Calado, A., Dias, M. S., & Braga, D. (2012). Impact of age in ASR for the elderly: preliminary experiments in European Portuguese. In *Advances in speech and language technologies for Iberian languages* (pp. 139–147). Springer.

Pellom, B., & Hacioglu, K. (2001). *SONIC: The University of Colorado Continuous Speech Recognizer* (Tech. Rep. No. TR-CSLR-2001-01). Boulder, Colorado: University of Colorado.

Peri, R., Somandepalli, K., & Narayanan, S. (2022). *To train or not to train adversarially: A study of bias mitigation strategies for speaker recognition.* arXiv. Retrieved from https://arxiv.org/abs/2203.09122 doi: 10.48550/ARXIV.2203.09122

Pervaiz, A., Hussain, F., Israr, H., Tahir, M. A., Raja, F. R., Baloch, N. K., … Zikria, Y. B.

(2020). Incorporating noise robustness in speech command recognition by noise augmentation of training data. *Sensors*, *20*(8). Retrieved from `https://www.mdpi.com/1424-8220/20/8/2326` doi: 10.3390/s20082326

Philippe Geubels als Vlaamse stem van Google Assistent. (2021). *Het Belang van Limburg*. Retrieved from `https://www.hbvl.be/cnt/dmf20211006_92691948` (Accessed: 2022-08-14)

Povey, D., Ghoshal, A., Boulianne, G., Burget, L., Glembek, O., Goel, N., … Vesely, K. (2011). The Kaldi speech recognition toolkit. In *Proceedings of the 2011 IEEE workshop on automatic speech recognition and understanding.* IEEE Signal Processing Society.

Povey, D., Peddinti, V., Galvez, D., Ghahremani, P., Manohar, V., Na, X., … Khudanpur, S. (2016). Purely sequence-trained neural networks for ASR based on lattice-free MMI. In *Proceedings of Interspeech 2016* (pp. 2751–2755).

Priya, M. B., & Kannamal, E. (2020). Investigation of speech recognition system and its performance. In *Proceedings of the 2020 International Conference on Computer Communication and Informatics (ICCCI)* (pp. 1–4). doi: 10.1109/ICCCI48352.2020.9104068

Qian, Y.-m., Weng, C., Chang, X.-k., Wang, S., & Yu, D. (2018). Past review, current progress, and challenges ahead on the cocktail party problem. *Frontiers of Information Technology & Electronic Engineering*, *19*(1), 40–63. doi: https://doi.org/10.1631/FITEE.1700814

Rabiner, L., & Juang, B. (1986). An introduction to hidden Markov models. *IEEE ASSP Magazine*, *3*(1), 4–16. doi: 10.1109/MASSP.1986.1165342

Rabiner, L., & Juang, B.-H. (1993). *Fundamentals of speech recognition.* USA: Prentice-Hall, Inc.

Russell, M., & D'Arcy, S. (2007). Challenges for computer recognition of children's speech. In *Proceedings of Speech and Language Technology in Education (SLaTE 2007)* (pp. 108–111).

Samardžić, T., Scherrer, Y., & Glaser, E. (2016). ArchiMob - a corpus of spoken Swiss German. In N. Calzolari et al. (Eds.), . Paris, France: European Language Resources Association (ELRA). Retrieved from `https://archive-ouverte.unige.ch/unige:91722` (ID: unige:91722)

Sawalha, M., & Abushariah, M. A. M. (2013). 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.*

Schalkwyk, J., Beeferman, D., Beaufays, F., Byrne, B., Chelba, C., Cohen, M., … Strope, B. (2010). "your word is my command": Google search by voice: A case study. In A. Neustein (Ed.), *Advances in speech recognition: Mobile environments, call centers and clinics* (pp. 61–90). Boston, MA: Springer US. Retrieved from `https://doi.org/10.1007/978-1-4419-5951-5_4` doi: 10.1007/978-1-4419-5951-5_4

Schiel, F., Kipp, A., & Tillmann, H.-G. (1998). Statistical modelling of pronunciation: it's not the model, it's the data. In *Modeling pronunciation variation for automatic speech recognition* (pp. 131–136).

Schuppler, B., Adda-Decker, M., & Morales-Cordovilla, J. A. (2014). Pronunciation variation in read and conversational Austrian German. In *Proceedings of Interspeech 2014* (pp. 1453–1457). doi: 10.21437/Interspeech.2014-355

Schuppler, B., Ernestus, M., Scharenborg, O., & Boves, L. (2011). Acoustic reduction in conversational dutch: A quantitative analysis based on automatically generated segmental transcriptions. *Journal of Phonetics*, *39*(1), 96–109. Retrieved from `https://www.sciencedirect.com/science/article/pii/S0095447010000926` doi: https://doi.org/10.1016/j.wocn.2010.11.006

Schwartz, R., Colthurst, T., Duta, N., Gish, H., Iyer, R., Kao, C.-L., … Chen, L. (2004). Speech recognition in multiple languages and domains: the 2003 BBN/LIMSI EARS system. In *Proceedings of the 2004 IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)* (Vol. 3, pp. 753–756). doi: 10.1109/ICASSP.2004.1326654

Sikka, D. (n.d.). *Speech recognition and AI.* https://studentsxstudents.com/speech-recognition-and-ai-82fd74f72bf2. (Accessed: 22-08-01)

Silver, S. (2022). *A history of voice technology.* https://info.keylimeinteractive.com/history-of-voice-technology. (Accessed: 2022-08-01)

Sivaraman, G., Mitra, V., Nam, H., Tiede, M., & Espy-Wilson, C. (2016). Vocal tract length normalization for speaker independent acoustic-to-articulatory speech inversion. In *Proceedings of Interspeech 2016* (pp. 455–459). Retrieved from http://dx.doi.org/10.21437/Interspeech.2016-1399 doi: 10.21437/Interspeech.2016-1399

Snyder, D., Chen, G., & Povey, D. (2015). Musan: A music, speech, and noise corpus. *arXiv preprint arXiv:1510.08484*.

Sreeram, G., & Sinha, R. (2020). Exploration of end-to-end framework for code-switching speech recognition task: Challenges and enhancements. *IEEE Access*, *8*, 68146–68157. doi: 10.1109/ACCESS.2020.2986255

Steurs, F., Vandeghinste, V., & Daelemans, W. (2022, 2). *Report on the Dutch language* (Tech. Rep. No. D1.10). European Language Equality (ELE). Retrieved from https://european-language-equality.eu/wp-content/uploads/2022/03/ELE___Deliverable_D1_10__Language_Report_Dutch_.pdf

Stevens, S. S., Volkmann, J., & Newman, E. B. (1937). A scale for the measurement of the psychological magnitude pitch. *The Journal of the Acoustical Society of America*, *8*(3), 185–190.

Stolcke, A. (2002). SRILM - an extensible language modeling toolkit. In *Proceedings of the 7th International Conference on Spoken Language Processing (ICSLP 2002)* (pp. 901–904). doi: 10.21437/ICSLP.2002-303

Sun, T., Gaut, A., Tang, S., Huang, Y., ElSherief, M., Zhao, J., ... Wang, W. Y. (2019). *Mitigating gender bias in natural language processing: Literature review.* arXiv. Retrieved from https://arxiv.org/abs/1906.08976 doi: 10.48550/ARXIV.1906.08976

Taeldeman, J. (2001). De regenboog van de Vlaamse dialecten. In J. Taeldeman, M. Devos, & J. De Caluwe (Eds.), *Het taallandschap in Vlaanderen* (pp. 49–58). Academia Press.

Tatman, R. (2017). Gender and dialect bias in YouTube's automatic captions. In *Proceedings of the first ACL workshop on ethics in natural language processing* (pp. 53–59).

Tatman, R., & Kasten, C. (2017). Effects of talker dialect, gender & race on accuracy of Bing Speech and YouTube automatic captions. In *Proceedings of Interspeech 2017* (pp. 934–938). doi: 10.21437/Interspeech.2017-1746

Technologies, S. L. (2021). *Language support in voice assistants compared.* https://summalinguae.com/language-technology/language-support-voice-assistants-compared/. (Accessed: 2022-08-09)

Toussaint Hutiri, W., & Ding, A. Y. (2022). Bias in automated speaker recognition. In *Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency* (pp. 230–247). New York, NY, USA: Association for Computing Machinery. Retrieved from https://doi.org/10.1145/3531146.3533089 doi: 10.1145/3531146.3533089

Van Bezooijen, R., & Van den Berg, R. (1999a). Taalvariëteiten in Nederland en Vlaanderen: hoe staat het met hun verstaanbaarheid? *Taal en Tongval*, *51*, 15–33. Retrieved from https://hdl.handle.net/2066/132061

Van Bezooijen, R., & Van den Berg, R. (1999b). Word intelligibility of language varieties in the Netherlands and Flanders under minimal conditions. *Linguistics in the Netherlands*, *16*(1), 1–12.

Van Eerten, L. (2007). Over het Corpus Gesproken Nederlands. *Nederlandse Taalkunde*, *12*, 194–215.

Van Wijngaarden, S. J. (2001). Intelligibility of native and non-native Dutch speech. *Speech Communication*, *35*(1-2), 103–113.

Vipperla, R., Renals, S., & Frankel, J. (2008). Longitudinal study of ASR performance on ageing voices. In *Proceedings of Interspeech 2008* (pp. 2550–2553).

Vipperla, R., Renals, S., & Frankel, J. (2010). Ageing voices: The effect of changes in voice parameters on ASR performance. *EURASIP Journal on Audio, Speech, and Music Processing*, *2010*, 1–10.

Waibel, A., Hanazawa, T., Hinton, G., Shikano, K., & Lang, K. (1989). Phoneme recognition using time-delay neural networks. *IEEE Transactions on Acoustics, Speech, and Signal Processing*, *37*(3), 328–339. doi: 10.1109/29.21701

Wang, D., & Zheng, T. F. (2015). Transfer learning for speech and language processing. In *Proceedings of the 2015 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA)* (pp. 1225–1237). doi: 10.1109/APSIPA.2015.7415532

Wang, J., Zhu, Y., Fan, R., Chu, W., & Alwan, A. (2021). Low Resource German ASR with Untranscribed Data Spoken by Non-Native Children — INTERSPEECH 2021 Shared Task SPAPL System. In *Proceedings of Interspeech 2021* (pp. 1279–1283). doi: 10.21437/Interspeech.2021-1974

Watanabe, S., Hori, T., Karita, S., Hayashi, T., Nishitoba, J., Unno, Y., ... Ochiai, T. (2018). Espnet: End-to-end speech processing toolkit. In *Proceedings of Interspeech 2018* (pp. 2207–2211). Retrieved from `http://dx.doi.org/10.21437/Interspeech.2018-1456` doi: 10.21437/Interspeech.2018-1456

Wepner, S. (2021). Adaptation of automatic speech recognition systems to the needs of Austrian German.. (Phonetikworkshop 46. Österreichische Linguistiktagung 2020; Conference date: 04-12-2020 Through 06-12-2020)

Xiong, W., Droppo, J., Huang, X., Seide, F., Seltzer, M., Stolcke, A., ... Zweig, G. (2017). The Microsoft 2016 conversational speech recognition system. In *Proceedings of the 2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)* (pp. 5255–5259). doi: 10.1109/ICASSP.2017.7953159

Xu, H., Li, K., Wang, Y., Wang, J., Kang, S., Chen, X., ... Khudanpur, S. (2018). Neural network language modeling with letter-based features and importance sampling. In *Proceedings of the 2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)* (pp. 6109–6113).

Young, S., Kershaw, D., Odell, J., Ollason, D., Valtchev, V., & Woodland, P. (2000). *The HTK Book (for HTK Version 3.0).* Microsoft Corporation.

Zabel, S., & Otto, S. (2021). Bias in, bias out–the similarity-attraction effect between chatbot designers and users. In *International Conference on Human-Computer Interaction* (pp. 184–197).

Zhang. (2022). *Bias mitigation against non-native speakers in dutch asr* (Master's thesis, Delft University of Technology). Retrieved from `https://repository.tudelft.nl/islandora/object/uuid%3Adf87fbca-7e88-4ea8-858b-3b8f4a194c87`

Zhang. (2022). *Mitigating bias against non-native accents* (Master's thesis, Delft University of Technology). Retrieved from `http://resolver.tudelft.nl/uuid:bc989a6e-60b5-4cff-bb7f-999c616afc7c`

Zhou, V. (2019). *Machine learning for beginners: An introduction to neural networks.* `https://victorzhou.com/blog/intro-to-neural-networks/`. (Accessed: 2022-08-02)

Zumalt, J. R. (2005, Dec.). Voice recognition technology: Has it come of age? *Information Technology and Libraries*, *24*(4), 180–185. Retrieved from `https://ejournals.bc.edu/index.php/ital/article/view/3382` doi: 10.6017/ital.v24i4.3382