Comparing performance of ASR systems on native Dutch children and teenagers: Google vs. Microsoft

Evaluating Speech Recognition Accuracy of state-of-the-art ASR models on Dutch child and teenager speech

Gert van Dijk[1]

Supervisor(s): Odette Scharenborg[1]**, YuanYuan Zhang**[1]

[1]EEMCS, Delft University of Technology, The Netherlands

A Thesis Submitted to EEMCS Faculty Delft University of Technology,
In Partial Fulfilment of the Requirements
For the Bachelor of Computer Science and Engineering
June 23, 2024

Name of the student: Gert van Dijk
Final project course: CSE3000 Research Project
Thesis committee: Odette Scharenborg, YuanYuan Zhang, Catharine Oertel

Abstract

Automatic Speech Recognition (ASR) technology is becoming more and more useful in everyday life, therefor also requiring higher accuracy across all different user demographics. This study compares the performance of Google's and Microsoft's ASR systems on native Dutch child and teenager speech using the JASMIN-CGN dataset as ASR for children presents unique challenges due to their shorter vocal tracts and irregular speech patterns. This research evaluates each system's performance based on Word Error Rate (WER) and Character Error Rate (CER), highlighting the differences between gender, age, and dialect regions. The results indicate that while Microsoft's ASR consistently outperforms Google's in terms of WER, Google demonstrates slightly higher precision in terms of CER. Therefor Microsoft is considered the better overall performing system but depending on one's needs, such as precision, Google would be the more favorable one.

Index Terms: speech recognition, Child speech recognition, ASR, Google, Microsoft, JASMIN-CGN, Dutch, teenagers, children

1. Introduction

1.1. Introduction

Automatic speech recognition (ASR) is the technology that converts spoken human language into text. This allows for fast and more natural communication between humans and machines, as speech is often seen as the most natural form of communication. Despite showing great improvements over the past few years, ASR technology still suffers from biases which can stem from varying factors such as gender, age, speech patterns, nationality, and even certain medical conditions. Some recent studies, such as [1], have tried to look deeper into these biases, revealing, for example, that female speech is recognised better than male speech [2] [3]. Feng et al. [2] also found that there was a bias depending on the age of the speaker as people under the age of 30 were reported to be better recognized than people above the age of 30.

There can be a multitude of reasons for the differences in performance between specific groups of speakers in ASR systems, such as biases in the training data, a lack of good data but also simply the difficulty of some speech groups. Child speech recognition (CSR) is a great example of the latter. Child speech is more difficult for ASR than most other age groups due to children having shorter vocal tracts and their more irregular speech patterns, stemming from their inconsistent speed of speech and their pronunciation. The limited accessible data of child speech further complicates this issue [3] [4] [5] [6] as all ASR systems require a good amount of training data to be able to perform. Next to that it is easier for models to find patterns when they exist making child speech trickier than some other groups due to their inconsistencies. Some studies tried to find a solution for the lack of data by using data augmentation on a small data set. Each study [4] [5] [6] came to the conclusion that by using some type of data augmentation, each using a different one, the performance of ASR on child speech improved, despite the lack of sufficient data.

This paper aims to examine the performance of state-of-the-art ASR systems, specifically Google and Microsoft's, on native Dutch child speech using the JASMIN-CGN dataset. Being two of the worlds leading companies in technology that offer ASR systems it will be a good baseline for future work of how well the state-of-the-art ASR systems perform. It will also serve as a benchmark for the ASR industry itself, showing what is currently possible and what is still lacking. By specifically comparing the two with each other it will not only reveal which is the better one but also potentially stimulate improvements through competition. not only that but it can also be used as reference for users wanting to choose the most appropriate ASR system that fits their requirements. Lastly, since both Google and Microsoft are so large, their user base will also be huge meaning this paper will have more relevance to more people than some other smaller ASR systems as many people might never interact with those services unlike with Google and Microsoft.

For the experiments themselves the JASMIN-CGN dataset was used, that released in 2004, containing important demographics such as children, elderly people and non-native speakers [7]. Very little research so far has been done on the Dutch language, some examples being [1] and [8], and in particularly for underrepresented groups, therefor this study attempts to better map the current state of both native child and teenage speech. By focusing on those two groups in particular, it will become more clear if the current systems are performing better or worse on these groups compared to the other groups. Not only that but children also make up a good part of the people that use these systems and therefor it is important that they function properly. Lastly by improving the ASR systems for child speech, or at the very least knowing potential areas where to improve on, it can be used more effectively in things such as education, accessibility features, and even interactive application aimed at younger audiences

In summary, this research will compare and analyze the current performance of two ASR systems, Google and Microsoft, on native Dutch child and teenager speech and propose potential future works. The following sections will go over the methodology, the ASR systems themselves and their backgrounds, the results of the experiment and potential future works.

1.2. Problem description

The main question of this research is "How do Google and Microsoft's ASR API compare when ran on native Dutch child and teenager speech". It will aim to analyze how two of the leading companies in technology that offer an ASR service, Google and Microsoft, perform on an underrepresented group of speakers.

By splitting the main research question up into smaller subquestions it will become easier to answer it. This will be done by comparing different splits, such as gender and dialect region, within the child and teenager speech files. The first group was the difference between male and female speakers. This was chosen since there have been papers that stated that there was a difference in results when comparing male and female speech, so it is important to verify whether this also applies to child and teenager speech or not and to see if Google and Microsoft also display this difference. Since there are different accents within the Netherlands it was important to see if there was any difference in performance based on what regional accent the speaker has. This lead to the following subquestions:

- How well does Google's ASR API perform on native Dutch child and teenager speech in terms of WER and CER?
- How well does Microsoft's ASR API perform on native Dutch child and teenager speech in terms of WER and CER?
- What differences can be observed between the performance on native Dutch male and female children and teenagers, when using both Microsoft's ASR API and Google's ASR API?

- What differences can be observed between Google and Microsoft's ASR APIs performance on native Dutch child and teenager speech from different dialect regions?

2. Methodology

The research consisted of two key components, the publicly available Microsoft and Google ASR APIs and the JASMIN-CGN dataset. By preprocessing the data from JASMIN and feeding it into both Google and Microsoft it was possible to calculate the WER and CER for each speech file and represent it visually using tables and boxplots respectively. Based on these results further conclusions were made relative to the main research question stated previously.

2.1. The JASMIN-CGN dataset

The provided JASMIN-CGN dataset comes with pregrouped and ordered files. Since not all files are relevant for this paper only those that were used will be mentioned.

2.1.1. Gender

The dataset contains two genders, male and female, and there are a total of 28493 audio segments of which 13845 are female and 14648 are male recordings.

2.1.2. Age Group

Only two of the provided age groups were used in this research:

- **Group 1:** Children aged between 7 and 11 years old.
- **Group 2:** Teenagers aged between 12 and 16 years old.

There are 16826 audio segments of children and there are 11667 audio segments of teenagers.

2.1.3. Dialect Region

Next the available dialect regions that were used are:

- **N1b:** North-Holland, excluding West Friesland.
- **N2c:** Gelders river area, including Arnhem and Nijmegen.
- **N3b:** Overijssel.
- **N4a:** Noord-Brabant.

N1b has 4031 audio segments, N2c has 7936 audio segments, N3b has 9021 audio segments and N4a has 7505 audio segments.

2.2. Evaluation metrics

WER will be the main metric used to calculate the performance. WER indicates the percentage of words that were incorrect compared to the total amount of words. it is calculated as followed:

$$\text{WER} = \frac{S + D + I}{N} \times 100\% \tag{1}$$

$$\tag{2}$$

where S is the number of substitutions, D is the number of deletions, I is the number of insertions, and N is the number of words in the reference.

CER, Character Error Rate, will be used together with WER to further look at the performance of both systems. To calculate the CER this formula is used:

$$\text{CER} = \frac{S + D + I}{N} \times 100\% \tag{3}$$

$$\tag{4}$$

where S is the number of substitutions, D is the number of deletions, I is the number of insertions, and N is the number of characters in the reference.

Both of these metrics were chosen as they are commonly used by previously mentioned researches and is something discrete that can directly be compared with one another

2.3. The ASR Models

Two of the largest publicly available ASR models are developed by Google and Microsoft. These models achieve high accuracy in speech recognition through the use of advanced machine learning techniques such as deep learning.

2.3.1. Google ASR

Google's ASR system, also known as Google Cloud speech-to-text, is an ASR system build by Google themselves and is publicly available to anyone. Through the use of recurrent neural networks (RNNs) and more recently, transformer-based models, it is able to process audio data and transcribe it. Google trained their ASR system on large amounts of data collected mainly from their own products such as Google Voice Search and YouTube videos. This huge variety in data allows the models to be very flexible and learn from many different groups of speakers with different accents and languages.

2.3.2. Microsoft ASR

Microsoft's ASR system, also known as Azure speech-to-text, is an ASR system build by Microsoft themselves and is publicly available to anyone. Similar to Google's ASR, Microsoft's system uses deep neural networks, including convolutional neural networks (CNNs) and transformers, to achieve high-performance speech recognition. Microsoft also gets their data from their own products such as Xbox, Skype and Cortana which allows it to have a wide range of audio data, resulting in being able to train robust models.

Both Google and Microsoft use very similar state-of-the-art techniques for their ASR systems, ensuring high accuracy and robustness. Both systems are designed with the idea of covering a large area of unique and different speakers however this may not be as true for all groups as differences between performances on speakers have been reported and written about by various papers mentioned before.

2.4. Analysis method

This research will be comparing Google and Microsoft with one another. The comparison will be based on their performance on native Dutch child and teenager speech. Within this group some more comparison will be made. The first is the comparison between genders. Here the differences in performance on men and women will be looked at and how each system handles them. Next up a comparison will be made based on age. This means comparing child speech performance with teenager speech performance. Lastly there will be taken a look at the four available dialect regions, N1b, N2c, N3b and N4a, to see if any region outperforms the others. Putting these together results in the final comparison and loops back to the main research question of

how Google and Microsoft compare to one another in terms of WER and CER on native Dutch child and teenager speech.

2.5. Experiment

The following steps were taken to perform the experiments in this research paper. After having set up the dataframes in Python all that was needed was to setup a Google and Microsoft account to gain access to the APIs which is something both platforms guide the user through. Next a script was written that would allow to connect to either APIs and then sequence all speech segments one by one through the APIs, which can be done at the same time for both systems. These results were then written to their corresponding csv file. After running all the segments, a cleanup of the data was performed by ensuring the dataframes automatically ignored all segments that had 'ggg' in their groundtruth since these samples contain noises such as coughing that were deemed not important as this research is only interested in the performance on speech, and noises such as coughing are not considered human interpetable language.

3. Contribution

This research will contribute to the field of ASR systems regarding the Dutch language. It will try to address a notable gap in ASR research by testing the performances of ASR on a more under represented group of speakers, in this case native Dutch children and teenagers. Previous research has mainly focused on adult speech, leaving a gap in understanding the overall performance of ASR on the Dutch language. By delving deeper into this demographic it can work as future reference for improvements on ASR.

By comparing two of the largest and widely available ASR systems, Google and Microsoft, it should provide a good overview of the current state of ASR systems when it comes to the Dutch native child and teenager speech. Not only that but by comparing the two it should become more clear in what areas, that being gender, child, teenager and the dialect regions, one performance better than the other.

Through the use of the JASMIN-CGN dataset it was possible to get the required child and teenage speech files and it allowed for a research with enough samples to be representative of the overall performance.

Splitting the child and teenager speech into gender, dialect region and age groups, which are children aged between 7 and 11 years and teenagers aged between 12 and 16 years, it gave more insight into general performance based on aspects different from just age alone.

Based on all the results and shortcomings, potential future work will be suggested as it is important to state what could have gone better and what was lacking in this research for future researchers.

4. Results

All results are summarized and visualized in tables 1, 2, 3 and images 1, 2. Table 1 gives the numerical results when splitting the speech files up into the two age groups, those being children and teenagers. Table 2 shows the numerical results when splitting the speech files up into the two genders and table 3 does the same except then for the dialect regions, those being N1b, N2c, N3b and N4a. Figure 1 shows the difference in the WER score of Google minus Microsoft and figure 2 shows the same but for the CER score.

4.1. Tables

Table 1: *Statistics by Group: Average WER and CER for Google and Microsoft. "C/T" indicates "child/teenager". "Avg" indicates the average score over all the speakers. "Go/Mi" indicates "Google/Microsoft". "T-B" indicates "TDNN-BLSTM" which is the model used by [1]*

	WER			CER		
	C	T	Avg	C	T	Avg
Go	31.55	22.37	26.96	**20.34**	**15.71**	18.02
Mi	**26.96**	**16.44**	21.70	21.89	17.21	19.55
T-B*	39.35	26.85	33.10	-	-	-

Table 1 shows the average WER and CER score of both Google and Microsoft for children and teenagers. An extra entry for TDNN-BLSTM (T-B) was included which was obtained from [1]. The table shows that Microsoft performs better than both Google and T-B for children and teenagers in regards to the WER score. Google on the other hand outperforms Microsoft on both ages when it comes to the CER score.

Table 2: *Statistics by Group and Gender: Average WER and CER for Google and Microsoft. "F/M" indicates "Female/Male". "Avg" indicates the average score over all the speakers. "Go/Mi" indicates "Google/Microsoft". "T-B" indicates "TDNN-BLSTM" which is the model used by [1].*

	WER			CER		
	F	M	Avg	F	M	Avg
Go	26.34	27.55	26.95	**17.32**	**18.71**	18.02
Mi	**21.07**	**22.30**	21.69	19.12	19.96	19.54
T-B	31.85	33.63	32.74	-	-	-

Table 2 shows the statistics for both genders. From this it shows again that Microsoft performs better on both women and men that Google and T-B when it comes to the WER score. At the same time Google still performs better than Microsoft for both women and men in terms of CER.

Table 3 presents the statistics sorted by dialect region. Consistent with the other results, Microsoft performs the best on each dialect when it comes to the WER score. Notably Microsoft also outperforms google on the CER for N3b making this the only category in which Microsoft has a better CER than Google.

4.2. Graphs

While tables are able to provide numerical insights into the performances of each category, they lack the ability to show the distribution of these scores. Therefor it was decided to include two boxplots that represent the difference in WER and CER score between Google and Microsoft.

Table 3: *Statistics by Dialect: Average WER and CER for Google and Microsoft. "N1b" indicates "North-Holland excluding West Friesland". "N2c" indicates "Gelders river area, including Arnhem and Nijmegen". "N3b" indicates "Overijssel". "N4a" indicates "Noord-Brabant". "Go/Mi" indicate "Google/Microsoft".*

	WER		CER	
Dialect	Go	Mi	Go	Mi
N1b	21.49	**16.08**	**14.65**	16.76
N2c	29.38	**25.07**	**18.76**	21.77
N3b	29.38	**23.98**	19.23	**19.14**
N4a	27.58	**22.02**	**19.22**	20.80

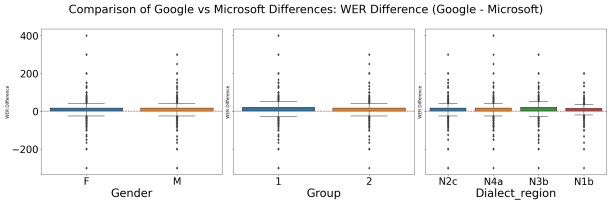

Figure 1: *Boxplots representing the difference of WER between Google and Microsoft for the 2 groups, genders and dialect regions.*

4.2.1. Genders

In figure 1 it can be observed that the interquartile range (IRQ) for men is wider than that of women indicating more inconsistencies in the WER differences between the two systems for men compared to women. In other words on average both Google and Microsoft have more consistent results for women than for men showing that there are potentially more uncertainties when it comes to male speech.

Figure 2, which looks at the CER, shows similar results to 1 however it can be seen that both Google and Microsoft experience a larger range and number of outliers.

4.2.2. Children and teenagers

Figure 1 shows that both Google and Microsoft are a lot more inconsistent for children than for teenagers seen by the large IRQ and median when compared to those of teenagers.

In figure 2 there is an even larger discrepancy between the IRQ and median of teenagers compared to that of children as for teenagers both the median and IRQ are even smaller and more condensed where as for children the IRQ and median got larger. Both figures show that Google and Microsoft seem to perform equally well relative to their respective results. That is to say if Microsoft does worse on A then on B then Google also does worse on A then on B but Microsoft could still do better on A then Google does on A.

4.2.3. Dialect regions

Both figure 1 and figure 2 show that in terms of the CER and WER both Google and Microsoft perform better on N1b than any other region. It shows that they seem to perform equally well relative to their respective results.

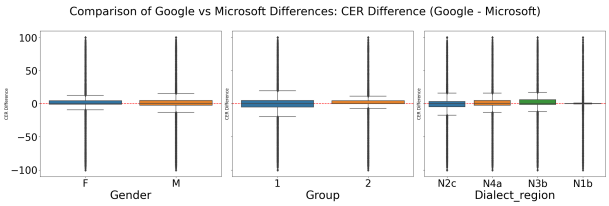

Figure 2: *Boxplots representing the difference of CER between Google and Microsoft for the 2 groups, genders and dialect regions.*

5. Responsible research

The JASMIN-CGN dataset has strict rules on who can use it and how it can be used. As it consists of speech files from children, elderly people and non native speakers it is important that their privacy is respected and uphold. As such no speech files were stored on a third-party storage such as Google and Microsoft. Microsoft states that unless specified by the user it shall not store any audio file send through their API on any of their servers. Google explicitly mentions that only files above 60 seconds will be stored, thus to ensure this did not happen all audio files were segmented into individual parts that were each below the 60 seconds mark.

Furthermore, this research is mainly focused on investigating potential bias within speech recognition systems. It is crucial to ensure that the technologies we develop and analyze do not have any existing biases against any group, especially vulnerable populations such as children, elderly people, and non-native speakers. By analyzing the performance of both Google and Microsoft's speech recognition systems across child and teenager speech, this research seeks to identify potential biases.

Lastly, All data was solely used for research purposes and no unethical use of data has occurred. For the dataset itself it has been stated that all audio recordings have been collected from people with informed consent. This means that all individuals that are present within the dataset were aware of the the uses of the recordings and gave their consent to be used in research. To ensure transparency on the actions taken with the dataset within this project, the methodology attempts to describe all performed actions as detailed as possible. All data stayed anonymous and no specific speaker was ever named even when looking directly at certain spoken segments. By following all these restrictions, this papers aims to uphold all the ethical standards and accepted rules set by the JASMIN-CGN dataset.

6. Discussion

The experiments in this study were conducted to evaluate the performance of Google and Microsoft's ASR APIs on native Dutch child and teenager speech through the usage of the Word Error Rate (WER) and the Character Error Rate (CER) as metrics. The results below are meant to portray the differences found between the two systems based on the previously mentioned and measured statistics.

6.1. Gender analysis

Overall the results show that Microsoft outperforms Google in terms of WER. Google slightly out performs Microsoft in terms

of CER yet this margin was so small its nearly neglectable. When comparing men and women directly against each other it can be seen that for both Microsoft and Google women are recognized better than men.

6.2. Child and teenager analysis

Similar to the results for gender, Microsoft out performs Google in terms of WER yet Google seems to do slightly better than Microsoft in terms of CER. When directly comparing children and teenagers it shows that teenagers are recognized significantly better than children in terms of WER and CER with a near 10 percent gap for the WER and a 5 percent gap for the CER. Although teenagers did have a significantly smaller sample size it should still be large enough to make for a fair comparison between teenagers and children.

6.3. Dialect analysis

Again Google underperforms compared to Microsoft in terms of WER yet performs better in terms of CER. As for the performance between the different dialects it appears that N1b is the best recognized one for both systems in terms of WER and CER. Looking at the other three regions it seems they are all relatively equally recognized with little to no variance between them. These results are consistent across both Google and Microsoft.

6.4. implications

For gender, child, teenager and each dialect Microsoft scores better in terms of WER than Google where as Google scores better in terms of CER however what does this actually imply. To have a lower CER could indicate that there are fewer substitution, deletion or insertion errors at the character level implying that Google might be slightly better at capturing finer details. Conversely to have a lower WER it could indicate that Microsoft is better at recognizing whole words and the correct sequence of words. Both have different use cases where it is more important than the other. Google having a better CER is most likely more favored for tasks that require precision such as transcribing names or singular words. Microsoft however might be more preferred when the understanding of the overall meaning of the sentence is important. An example of this would be for voice commands.

6.5. Short comings

Since this research was limited to 10 weeks it had it's fair share of shortcomings. The first was the manual checking of results to find more specific patterns in the common mistakes the systems made. Doing this would have provided with more in depth results highlighting more correctly what areas the systems were actually failing at. Since the WER and CER can show the overall performance however not the individual words or sentences it struggles with. Following up on the shortcomings of WER and CER, some more metrics could have been used to go into more detail onto some aspects that WER and CER do not show. One example would be the Phoneme Error Rate (PER) which provides insight into the system's ability to recognize phonemes correctly, something both CER and WER can not do. Lastly, a larger dataset would have allowed for a better and more comprehensive analysis. The JASMIN dataset does not cover all dialect regions, as it only contained four out of the sixteen dialect regions for child and teenager speech, which makes this analysis not accurate when talking about the entire Dutch language. Next to that there were some imbalanced comparisons in terms of number of audio segments. An example was children and teenagers with there being 16826 audio segments for children but only 11667 for teenagers.

7. Conclusions and future work

This study attempted to compare the performances between the Google and Microsoft ASR APIs regarding native Dutch child and teenager speech. It showed that Microsoft scored on average better than Google. In terms of WER Microsoft scored lower than Google for every category, those being gender, age and dialect, and despite performing less in terms of CER than Google it is still at all times within a 2 percent margin compared to the average 5 percent difference in the WER.

For gender both systems displayed an improved performance when it came to women compared to men. Microsoft overall however performed better on both men and women than Google did, showing a more consistent performance across both genders. This difference however is merely 1 to 2 percent indicating it is most likely not a bias created from the models themselves but from the overall complexity of understanding women compared to men.

Looking at both child and teenager speech it was found that both ASR systems struggle more on child speech then on teenager speech. Both systems showed a noticeable gap in terms of WER and CER having a roughly 10 percent difference in WER and 5 percent difference in CER. This indicates that both systems are less capable of recognising the overall structure of the sentences of children compared to those of teenagers. Not only that but it also struggles more when it comes to correctly recognizing individual characters. This would put younger people at a disadvantage when using systems that are voice assisted or even voice reliant as they will have a harder time to be understood correctly therefor making it more difficult to properly use them.

Delving into the dialects it was seen that N1b was recognized the best in terms of WER and CER for both Google and Microsoft. This difference in performance might be caused by the fact that N1b has the most speakers compared to the other groups and also is home of the capital city Amsterdam. This would make this group of speakers more appealing and possibly important for big companies like Google and Microsoft. This however does make it so that people that live outside of the capital will have a harder time using these systems.

To wrap it all up, looking at the overall performance of Google and Microsoft it seems that Microsoft is better at recognizing the full words and the sequence of words in a sentence. Google on the other hand seemed to be more precise as shown by the lower average CER score. Taking both factors into account Microsoft does perform on average slightly better than Google but depending on the use case one might still chose Google over Microsoft, those being mostly precision focused applications.

Being limited to a 10 week period this research had to cut corners in some areas as mentioned in the shortcomings. The main improvements that could be made would be a more in depth analysis on the results by going over the individual mistakes and trying to find patterns within the words that were often transcribed incorrectly. Next by adding additional metrics such as PER to get a further understanding of the performances of both systems since WER and CER both have their shortcomings such as not being able to represent the

phonemes. Lastly by having access to more data it would allow for a better overall research as the JASMIN dataset only cover four out of the sixteen dialect regions in the Netherlands.

8. Acknowledgements

I will thank Thomas, my team member, for sharing his code with me for segmenting all the audio files.

9. References

[1] S. Feng, O. Kudina, B. M. Halpern, and O. Scharenborg, "Quantifying bias in automatic speech recognition," 3 2021. [Online]. Available: https://arxiv.org/abs/2103.15122

[2] M. Sawalha and M. A. Shariah, "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," 2013. [Online]. Available: https://www.semanticscholar.org/paper/The-effects-of-speakers'-gender%2C-age%2C-and-region-on-Sawalha-Shariah/0caa84f35a586e7f91fb014566642bc943cbf83b

[3] M. Adda-Decker and L. Lamel, "Do speech recognizers prefer female speakers?" in *Interspeech*, 2005. [Online]. Available: https://api.semanticscholar.org/CorpusID:8668656

[4] S. Shahnawazuddin, N. Adiga, K. Kumar, A. Poddar, and W. Ahmad, "Voice conversion based data augmentation to improve children's speech recognition in limited data scenario." in *Interspeech*, 2020, pp. 4382–4386.

[5] G. Chen, X. Na, Y. Wang, Z. Yan, J. Zhang, S. Ma, and Y. Wang, "Data augmentation for children's speech recognition–the" ethiopian" system for the slt 2021 children speech recognition challenge," *arXiv preprint arXiv:2011.04547*, 2020.

[6] T. Patel and O. Scharenborg, "Improving end-to-end models for children's speech recognition," *Applied Sciences*, vol. 14, no. 6, p. 2353, 2024.

[7] C. Cucchiarini, H. Van Hamme, O. Van Herwijnen, and F. Smits, "JASMIN-CGN: Extension of the Spoken Dutch Corpus with Speech of Elderly People, Children and Non-natives..." *ResearchGate*, 1 2006. [Online]. Available: https://www.researchgate.net/publication/237245075_JASMIN-CGN_Extension_of_the_Spoken_Dutch_Corpus_with_Speech_of_Elderly_People_Children_and_Non-natives_in_the_Human-Machine_Interaction_Modality

[8] S. Feng, B. M. Halpern, O. Kudina, and O. Scharenborg, "Towards inclusive automatic speech recognition," *Computer speech language*, vol. 84, p. 101567, 3 2024. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0885230823000864