Universiteit Utrecht

Measuring the performance of an automatic speech recognition system: The effect of speaker gender and speech register.

Bachelor thesis Artificial Intelligence
Utrecht University
7,5 ECTS
Iris Leliveld
6259324

Supervisor: Frans Adriaans
Second reader: Bjørn Jespersen

June 26th, 2020

ABSTRACT

Speech recognition is an important part of artificial intelligence and has gotten a lot better over the years, but there is still room for improvement. Often the automatic speech recognition systems are not trained equally on male and female voices. Speech recognition systems have many uses, for example it can be used to research language acquisition by evaluating their performance on child directed speech. Child directed speech is a speech register that is used when speaking to children. When the two factors of speaker gender and speech register are combined, what does this mean for the performance of an automatic speech recognition system? In this thesis an answer will be given to the question "What is the effect of speaker gender and speech register on the performance of an automatic speech recognition system?". The output of an existing automatic speech recognition system was evaluated on accuracy, precision and recall, for a male and female voice using child directed speech and adult directed speech. It was found that speaker gender could positively influence performance if the system was trained on that gender. If that was not the case performance would be more negatively influenced. Furthermore, child directed speech has a positive influence on performance in comparison to adult directed speech.

Keywords: Child directed speech, automatic speech recognition system, gender, performance.

Chapter index

1. Introduction

Artificial intelligence is a field that is growing and is becoming more and more important, as can be seen in the amounts of money that are being invested in AI startups. In 2010 there was a total of $1.3B invested globally in AI startups, in 2018 this number has risen to $40.4B (Perault et al., 2019). "Funding has increased with an average annual growth rate of over 48% between 2010 and 2018" (Perault et al., 2019, p.88).

Artificial intelligence is used and can be used in a wide variety of fields. From self-driving cars to healthcare to home assistants. Speech recognition is an important part of artificial intelligence, because a lot of intelligent systems – like Siri and Alexa – use speech recognition.
Therefore it is important that speech recognition functions well on all kinds of voices and speech. The performance of automatic speech recognition systems keeps improving. This can be seen in the percentage of word error rates (WER), the number of errors (made by the system) divided by the number of words. A low WER percentage means that the system is pretty accurate, a high WER percentage means that the system is not very accurate (Rev, n.d.). Where in 2007 it wasn't uncommon to have a system with a WER around 27% (Sha & Saul, 2007). Ten years later this has dropped to 5,1% for a automatic speech recognition system of Microsoft (Xiong et al. 2018).

Currently however, there seems to be a gender bias in speech recognition where male voices are recognized better than female voices (Tatman, 2017; Rodger & Pendharkar, 2004). Tatman (2017) showed that the WER of youtube's automatic captions was significantly higher for women than it was for men. The mean WER for women was around 50%, while the mean WER for men was around 38%. Tatman (2017) explained these differences by discussing the data the system is trained on. The data in the speech corpora that are used to train these systems generally consist of more male speakers than female speakers. There are differences between male and female voices (further discussed in chapter 2.2) and because of these differences an automatic speech recognition system will perform worse on female voices when mostly trained on male voices.

Speech recognition systems can also help us give an insight into language acquisition, by evaluating its performance on child directed speech. Child directed speech is a special speech register, i.e. way of speaking, that people use when talking to infants/children. Child directed speech differs from adult directed speech, speech aimed at an adult, in multiple ways. Children prefer child directed speech (Cooper & Aslin, 1990) and most importantly, it is likely that child directed speech facilitates language acquisition. By combining speech recognition and child directed speech it can further our

understanding of how language acquisition works, because the speech recognition system can serve as a model for how language acquisition works in children.
Typically when researching child directed speech, it is mostly focused on mothers their use of child directed speech (Broesch & Bryant, 2018). There is less research about fathers use of child directed speech. This results in the fact that when speech recognition systems are trained on child directed speech to research language acquisition they are trained on data that consists mostly of female speakers (Kirchhoff & Schimmel, 2018; Ludusan, Cristia, Martin, Mazuka, & Dupoux, 2016). Because of the differences between male and female voices, one could hypothesize that these systems trained on primarily female voices, are not as good in recognizing male voices. So while these systems might have a problem with male voices, other systems have a problem with female voices.

Thus, when combining the factors speaker gender and speech register, what can an automatic speech recognition system recognize best? In this thesis an existing speech recognition system will be evaluated on the measures accuracy, precision and recall, to try and answer the research question "What is the effect of speaker gender and speech register on the performance of an automatic speech recognition system?" To answer this question an answer will be given to two sub-questions, "Is the automatic speech recognition system better at recognizing a male voice or a female voice?" and "Is the automatic speech recognition system better at recognizing child directed speech or is it better at recognizing adult directed speech?".

1.1. Expectations

There are different hypotheses made in regard to the two sub research questions. These hypotheses are based on the literature discussed in chapter 2. First the hypotheses for the first sub-question are given, then for the second sub-question.

In regards to the first sub-question there are two possible hypotheses. The first hypothesis is that male and female voices are recognized equally well, because the system was trained on both of them equally. The second hypothesis is that a male voice is recognized better than a female voice, because the system was trained more on male voices.

In regard to the second sub-question there are also two possible hypotheses.
The first hypothesis is that the system will recognize adult directed speech better than child directed speech, because it was trained on adult directed speech, but also because the pitch of child directed speech is higher. Kirchhoff and Schimmel (2005) found that when a system was trained on a certain speech register it was also better at recognizing this register. Furthermore it is expected that the system will perform worse when pitch is higher, because if the system is mostly trained on male

voices it is worse at recognizing female voices (Tatman, 2017; Rodger & Pendharkar, 2004). This difference in performance is caused by a difference between male and female voices, the most important difference between these two is pitch. So automatic speech recognition systems perform worse on female voices when trained on male voices, because female voices have a higher pitch. Thus, it will also perform worse on child directed speech.
A second hypothesis is that the system is better at recognizing child directed speech, because child directed speech is slower and more exaggerated than adult directed speech, which will make it easier to recognize. Furthermore child directed speech is easier to segment for children (Thiessen et al., 2005), so this could also be the case for a computer program.

1.2. Structure

The thesis is structured as follows: chapter 2 will discuss the theoretical background relevant to this thesis and is divided in subsections. Section 2.1 will discuss child directed speech and its characteristics. Section 2.2 will discuss the importance of child directed speech and section 2.3 will discuss the differences between male and female voices. Chapter 3 will discuss the method that is used to research the research question. Chapter 4 will discuss the results, in chapter 5 a discussion will be held and in chapter 6 a conclusion will be drawn. The used references can be found in chapter 7 and additional information, such as the transcriptions of the audio can be found in the appendix in chapter 8.

2. Theoretical background

In this chapter theoretical information will be discussed that is relevant to this thesis. Section 1 will discuss child directed speech and what its characteristics are that make it different from adult directed speech. These characteristics are the reason that there would be a difference in performance between adult directed speech and child directed speech.
Section 2 will discuss why child directed speech is important. Section 3 will discuss the differences between male and female voices. These differences are the reason why there is a difference in performance between male and female voices.

2.1. Characteristics of child directed speech

Child directed speech, also known as infant directed speech or motherese, is speech that is specifically aimed at children or infants. It is a different kind of speech than speech aimed at adults. Speech aimed at adults is known as adult directed speech. There are a few characteristics that distinguish child directed speech from adult directed speech.

The first and most important characteristic is that child directed speech has a higher fundamental frequency i.e. pitch (Fernald & Simon, 1984; Fernald et al., 1989; Grieser & Kuhl, 1988). Fernald et al. (1989) showed that this is most likely an universal characteristic. This was shown by analysing the speech from parents from different countries (France, Italy, Germany, Japan, UK and USA). They selected fifty utterances from parents' their child directed speech and fifty from their adult directed speech. The mean fundamental frequency differed per language. So the mean pitch during adult directed speech and child directed speech was higher in French than in German. They concluded that both mothers and fathers spoke with a higher fundamental frequency in their utterances directed at children. This feature was also found in the child directed speech of Chinese mothers who only spoke Mandarin (Grieser & Kuhl, 1988). Mandarin is a tonal language, this means that a change in pitch can change the meaning of a word. A quarter of the world's languages is tonal, so the presence of this feature in tonal languages contributes evidence for the hypothesis that this is an universal feature (Grieser & Kuhl, 1988).

The second characteristic is that child directed speech is more exaggerated than adult directed speech, e.g. more exaggerated vowels (Kuhl et al., 1997) and intonation (Fernald & Simon, 1984; Fernald & Mazzie, 1991). Kuhl et al. (1997) showed that American, Russian and Swedish mothers all used more acoustically extreme /i/, /a/, and /u/ vowels when speaking utterances directed at children. Fernald and Simon (1984) compared child directed speech from German mothers and found that 77% of child directed utterances contained prosodic patterns that are only rarely found in adult directed speech.

“These prosodic patterns consisted either of "expanded" pitch contours or whispered speech.” (Fernald & Simon, 1984, p.108). When mothers speak to children they use distinctive prosodic patterns to highlight focused words (Fernald & Mazzie, 1991).

The third characteristic is that child directed speech is slower, the phrases are shorter and there are longer pauses between words (Fernald et al., 1989; Fernald & Simon, 1984; Grieser & Kuhl, 1988). Grieser and Kuhl (1988) found that in speech that had a duration of fifteen minutes the average phrase duration was longer in adult directed speech (1.7 seconds) than in child directed speech (1.1 seconds). They also found that on average the pauses between words were longer in child directed speech (1.1 seconds) than in adult directed speech (0.8 seconds).

2.2. Importance of child directed speech

Children as young as two days already have a preference for infant directed speech over adult directed speech (Cooper & Aslin, 1990). Cooper and Aslin (1990) researched this by measuring how long a child looked at a grey panel while a recording of child directed speech played and while a recording of adult directed speech played. When the child was looking at the panel they had their attention on the speech and when they stopped looking they no longer had their attention on the speech. It was found that on average children looked at the panel for 18.3 seconds when a recording of child directed speech played. When a recording of adult directed speech played, the average looking time was 13.9 seconds. This is a significant difference (Cooper & Aslin, 1990). Cooper and Aslin (1990) found the same preference in children that were one month old.

So children have a preference for child directed speech, but why do they have this preference and why does it matter? There is evidence that suggests that child directed speech accommodates language development (Thiessen, Hill, & Saffran, 2005). Child directed speech helps by segmenting words (Thiessen et al., 2005) and learning phonetic categories (Werker et al., 2007).

When children learn language, a problem that they face is the word segmentation problem. Where does a certain word begin and where does it end? Child directed speech helps children with this problem by facilitating word segmentation (Thiessen et al., 2005). When children were exposed to words and parts of words in a nonsense language, they had a preference for words in child directed speech. In the adult directed speech on the other hand there was no preference for words over parts of words. Thus that children were able to distinguish between words and parts of words in child directed speech, but not in adult directed speech. This suggests that children find child directed speech easier to segment than adult directed speech (Thiessen et al., 2005).

Every language has a set of phonemes that are specific to that language. Phonemes are the smallest speech sounds and in writing they are represented by letters. For example the letter “a” is pronounced different in the words “face” and “bath”, so these are two different phonemes. For adults it is difficult to distinguish between different phonemes in a language that is not their native language. However very young children do not have this difficulty. They are able to distinguish between different phonemes, even if they are not in their native language (Werker, Gilbert, Humphrey, & Tees, 1981). As they get older they lose this ability and they are only able to distinguish between different phonemes that appear in their native language. So how does this happen? Maye, Werker and Gerken (2002) found that children are sensitive to the statistical distribution of phonemes in the input that they get. So when most of the input that a child gets is in English, it is the set of phonemes that appear in English that are deemed most important. Phonemes that appear in other languages are less important and will be disregarded. Child directed speech is relevant to this because it contains language specific cues for the phonetic categories of the native language (Werker et al., 2007).

2.3. Differences between male and female voices

There are certain differences between male and female voices. Here the differences in fundamental frequency and quality are discussed.

Between male and female voices there is a difference in fundamental frequency, e.g. pitch. This is caused by a difference in male and female vocal cords. Male vocal folds are generally longer and thicker than female vocal folds, this causes them to vibrate more slowly (Simpson, 2009). On average the fundamental frequency of German or English male speakers is 100 to 120 Hz (Simpson, 2009). Female vocal folds are shorter and lighter, so they vibrate at a higher frequency. They have a frequency around 200 to 220 Hz (Simpson, 2009). So male voices are perceived as lower pitched and female voices are perceived as higher pitched. It is important to note that the average pitch of a speaker is dependent on the language that they speak. For example, female speakers of Dutch have an average fundamental frequency of 189 Hz and male speakers of Dutch have an average fundamental frequency 111 Hz (Biemans, 1998). Whereas Rose (1991) found that female speakers of a Chinese dialect have an average fundamental frequency of 187 Hz and the male speakers of this dialect have an average fundamental frequency of 170 Hz.

Another difference in male and female voices is that female voices have a more “breathy” quality (Mendoza, Valencia, Muñoz, & Trujillo, 1996). Simpson (2009) explains breathy voice as follows: “During each cycle of normal vibration the vocal folds come together and briefly close the airway. In

breathy voice, however, the vocal folds do not close completely during the cycle, so that there is a constant flow of air through the glottis." p.623. Female vocal folds are thinner and this means that they never really close during each cycle, which allows air to escape (Titze, 1989). This causes female voices to have a more "breathy" quality than male voices.

3. Method

3.1. Participants

Due to time constraint and limitations due to COVID-19 it was decided to only use two participants. One participant was male and the other participant was female. The male participant was 50 years of age, the female participant was 20 years of age. The experimenter themself participated as the female participant. This was not deemed as a problem, because before the experiment was run the participants needed to familiarize themselves with the text. Thus it was not an issue that the female participant already knew which text would be used.

3.2. Materials

A Dutch piece of text was chosen from the website lingua.com. On this website texts in different languages can be found, on varying difficulty levels (A1 to B2). Both participants were native speakers of Dutch, thus the text "Ruzie in de supermarkt" was chosen. This was the only text on B2 level, which would come closest to the level of the native speakers. This text was also selected because an important characteristic of the text had to be that it could be told as if it was something the participant had experienced. That would make it easier for it to be told like it was something that the participant had experienced rather than just reading it aloud. This text was written from a first person perspective and talks about the ordinary event of going to the supermarket. The version found on lingua.com was slightly modified so the text would flow better when read aloud. The used text can be found in section 8.1 in the appendix.

The automatic speech recognition system that was used was a version of Kaldi and was implemented by Stichting Open Spraaktechnologie (Openspraaktechnologie, z.d.). An online version was used. This system was trained on the corpus "corpus gesproken Nederlands", a corpus of spoken Dutch, which was made available by Nederlandse Taalunie (Openspraaktechnologie, z.d.). This corpus was recorded between 1998 and 2004 and consists of both spoken Dutch and Flemish. It consists of different kinds of spoken language aimed at adults, like interviews and news stories. Hence why it is assumed that the system is trained on adult directed speech. It isn't clear what the ratio of male and female voices is (Lands.let.ru, z.d.).

Furthermore, the program PRAAT was used to analyse the pitch of the recordings and the duration. The audio was recorded using the recorder app on a Xiaomi mi 9t pro phone.

3.3. Procedure

First it was explained what child directed speech and adult directed speech was. Then the experiment started. The experiment consisted of two phases, a familiarization phase and a test phase.

In the familiarization phase the participant got three minutes to read through the text and to read the text aloud to themselves. The purpose of this phase was to familiarize the participant with the text. It was important that the participant knew what they would be reading aloud, to minimize reading mistakes and to avoid too much stumbling across the words, because this could be picked up by the automatic speech recognition system.

After the familiarization phase, the test phase started. In this phase there were two conditions. The adult directed speech (ADS) condition and the child directed speech (CDS) condition. In the adult directed speech condition the participant was asked to read the text aloud like they were telling it to an adult. In the child directed speech condition the participant was asked to read the text aloud like they were telling it to a child. Before recording happened it was explained that when reading the text aloud, it should be read like it was something the participant had just experienced yesterday. This was to ensure that it was spoken more as if it was an experience and not like they were just reading a text. So the goal of this was to simulate conversational speech. This was especially important in the adult directed speech condition, because when you read a text aloud you already speak very different then when you speak to another person.

This resulted in four audio recordings, a male voice using child directed speech and adult directed speech, and a female voice using child directed speech and adult directed speech. These recording were then given to the automatic speech recognition system and were labelled as “daily conversations”. This resulted in four text transcriptions. These transcriptions were then evaluated. The transcriptions can be found in section 8.2 in the appendix.

Furthermore, the mean pitch and the duration of the recordings was analysed used PRAAT.

3.4. Evaluation

Each transcription was evaluated on the following measures.

- Accuracy: of the amount of words that were actually said, how many are transcripted correctly?
- Precision: of the amount of words that the system transcripted, how many were actually said?
- Recall: of the amount of words that were actually said, how many were transcripted?

These measurements were chosen because they are often used when evaluating the performance of a system/model. The system isn't just evaluated on accuracy, because sometimes the accuracy can be really high while the system is not per se a good system (Developers.google, z.d.). That is why precision and recall were also chosen.

These measures are calculated using values from a confusion matrix.

		Predicted	
		Negative	**Positive**
Actual	**Negative**	True Negative	False Positive
	Positive	False Negative	True Positive

Figure 1: Confusion matrix (Ping Shung, 2018).

These values were interpreted as follows:

- True positives: the words the system transcripted that were actually said in the audio; the words that were correctly transcripted.
- False negatives, the words the system did not transcribe that were actually said in the audio; the words that should have been transcripted but weren't.
- False positives, the words the system transcripted that were not not said in the audio; the words that shouldn't have been transcripted.
- True negatives, the words the system did not transcribe that were not said in the audio. In this case the system does not predict words that weren't said, so this value is always zero.

All evaluations were done by hand and can be found in section 8.3 in the appendix. The following is an example of how evaluations were made. "We hadden al een tijd ruzie" is a part of the text, this part was transcripted by the system as: "badan tijd ruzie".

Input speaker	**Output system**	**True positives (cumulative)**	**False negatives (cumulative)**	**False positives (cumulative)**
We hadden al een	-	0	4	0
-	badan	0	4	1
tijd ruzie	tijd ruzie	2	4	1

Table 1: Evaluation example.

In some cases the system did correctly recognize a spoken word, however it was recognized as two separate words instead of one word. This was for example consistently the case for the word “ingaan” which was recognized as “in gaan”. These words were correctly recognized by the system and were thus counted as true positives, so “in gaan” is counted as one true positive. These words are coloured pink in the evaluations in the appendix. It also happened that during recording some words were added or that word order changed a little e.g. “dus besloten we maar om gewoon wat inkopen te gaan doen” became “dus besloten we om maar ~~om~~ gewoon wat inkopen te gaan doen”. These changes are indicated by a red coloured word and a crossed out word.

Once these evaluation values are known, accuracy, recall and precision were calculated using these formulas.

- Accuracy = (True positive + True negative) / (True positive + True negative + False positive + False negative) (Developers.google, z.d.), because here the true negative is always zero this formula can be simplified to
    - Accuracy = True positive / (True positive +False positive + False negative).
- Precision = True Positive / (True positive + False positive) (Ping Shung, 2018).
- Recall = True positive / (True positive + False negative) (Ping Shung, 2018).

4. Results

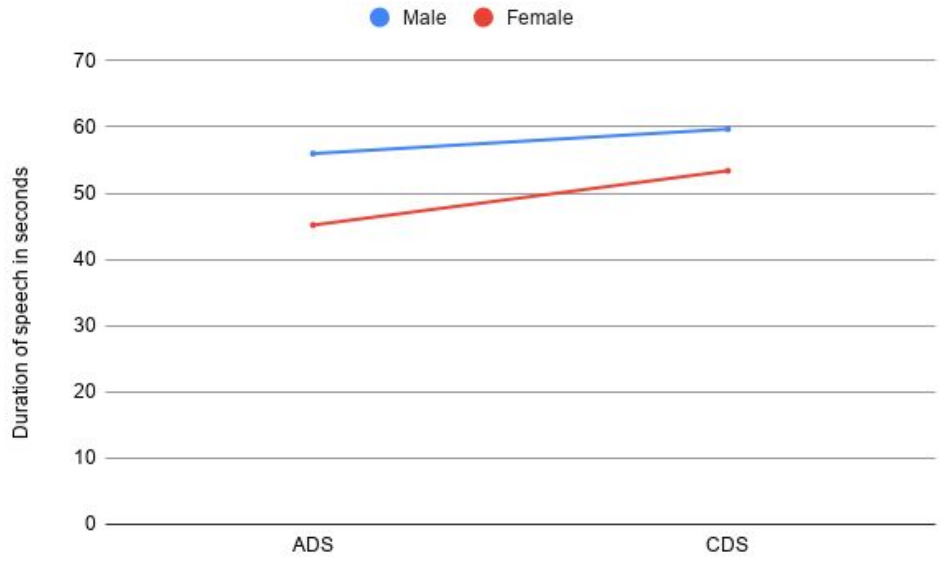
During recording of the audio some words were added to the text or the word order was changed a little. Hence why the total amount of spoken words is not exactly the same. There was also no significance test conducted. This was not possible, because of the small sample size.

	True positives	False negatives	False positives	Number of spoken words
Male ADS	136	32	15	168
Male CDS	151	16	6	167

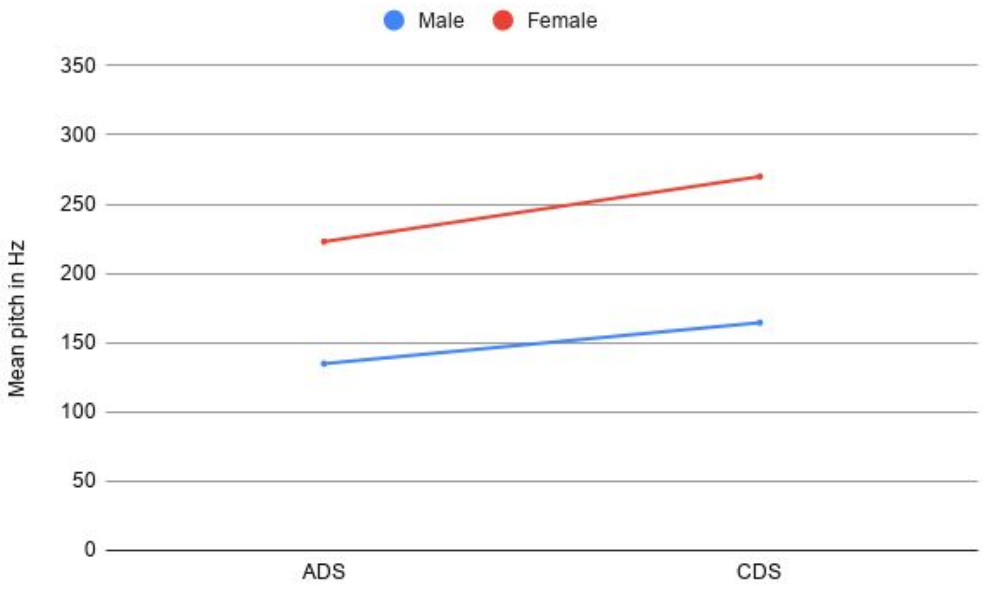
Table 2: Results male participant.

The male participant spoke 168 words in the ADS condition of which 136 words were true positives, 32 were false negatives and 15 were false positives.

In the CDS condition 167 words were spoken of which 151 words were true positives, 16 were false negatives and 6 were false positives. So when we compare these conditions, the amount of true positives increased and the amount of false negatives and false positives decreased. Mean pitch of a male speaker using adult directed speech is 134,9 Hz and this increases when using child directed speech to 164,5 Hz. Duration of speech also increased from 56,0 seconds in adult directed speech to 59,7 seconds in child directed speech.



Graph 1: Mean pitch of male and female speech during ADS and CDS condition.



Graph 2: Duration of male and female speech during ADS and CDS condition.

	True positives	False negatives	False positives	Number of spoken words
Female ADS	132	35	20	167
Female CDS	138	30	18	168

Table 3: Results female participant.

The female participant spoke 167 words in the ADS condition of which 132 words were true positives, 35 were false negatives and 20 were false positives.
In the CDS condition 168 words were spoken of which 138 words were true positives, 30 were false negatives and 18 were false positives. So when we compare these conditions, the amount of true positives increased and the amount of false negatives and false positives decreased. Mean pitch of a female speaker using adult directed speech is 223,0 Hz and this increases when using child directed speech to 270,0 Hz. Duration of speech also increased from 45,2 seconds in adult directed speech to 53,4 seconds in child directed speech.

When comparing these values between the male and female participant some differences can be noted. The amount of true positives increased more for the male participant in the CDS condition than for the female participant. In fact, the amount of true positives for the female participant in the CDS condition is almost the same as the amount of true positives for the male participant in the ADS condition. The amount of false negatives and false positives also decreased a lot more for the male

participant in the CDS condition than for the female participant. The amount of false negatives decreased by half for the male participant, while it only decreased slightly for the female participant. The amount of false positives also decreased a lot for the male participant, but it barely decreased for the female participant. In fact, the amount of false positives for the female participant in the CDS condition is higher than the amount of false positives for the male participant in the ADS condition. Furthermore both participants raised their pitch in the CDS condition by around 30 Hz and the duration of the speech was longer in the CDS condition.

	Accuracy	Precision	Recall
Male ADS	0.74	0.90	0.81
Male CDS	**0.87**	**0.96**	**0.90**
Female ADS	0.71	0.87	0.79
Female CDS	0.82	0.88	0.82

Table 4: Evaluation measures, best measures are bold.

The accuracy for a male speaker using adult directed speech is 0.74, yet when child directed speech is used this increases to 0.87. The accuracy for a female speaker using adult directed speech is 0.71, but when child directed speech is used this increases to 0.82. For both speakers the accuracy increases a lot when child directed speech is used, as opposed to adult directed speech. It increases with almost the same amount, however the accuracy for a female speaker is for both conditions a bit lower than for a male speaker.

The precision for a male speaker using adult directed speech is 0.90, yet when child directed speech is used this increases to 0.96. The accuracy for a female speaker using adult directed speech is 0.87, but when child directed speech is used this increases to 0.88.
So for both male and female speakers accuracy increases when child directed speech is used, however accuracy increases more for a male speaker and it barely increases for a female speaker. A female speaker using child directed speech has a lower precision than a male speaker using adult directed speech.

The recall for a male speaker using adult directed speech is 0.81, but when child directed speech is used this increases to 0.90. The recall for a female speaker using adult directed speech is 0.79, yet when child directed speech is used this increases to 0.82.

So for both male and female speakers recall increases when child directed speech is used, however recall increases more for a male speaker. In fact, the recall of a female speaker using child directed speech is almost the same as a male speaker using adult directed speech.

5. Discussion

5.1. Implications

In the results it was found that there is a difference in performance for male and female speakers, because of this we can assume that the system was not trained equally on both male and female speakers. So the found results one again emphasise the importance of training an automatic speech recognition system equally on both male and female voices. The results also show that even though the automatic speech recognition system is not specifically trained on child directed speech, it does perform better on this speech register than the speech register it was trained on: adult directed speech. This could mean that if you do train an automatic speech recognition system specifically on child directed speech, the performance of this system could increase significantly in comparison to an automatic speech recognition system trained on adult directed speech. However, this result of the system performing better on child directed speech when trained on adult directed speech is contradictory to the results found by Kirchhoff and Schimmel (2005). They found that a system trained on adult directed speech was also better at recognizing adult directed speech. Thus more research would have to be conducted on this subject to be able to draw a conclusion.

5.2. Limitations

There are a number of limitations that have to be considered when reading this thesis.

First of all the very small sample size has to be considered, since there were only two participants. This has serious consequences, because the results are not generalizable and it makes it very hard to verify if the differences between the evaluation values are significant or not. Furthermore the mean pitch of the speech of the female participant was very high. The mean pitch of the female participant was 223,0 Hz in the adult directed speech condition. The average mean pitch for female speaker of Dutch was 189 Hz as found by Biemans (1998). This means that in this research the pitch of the female speaker was 34 Hz higher than the average. Thus it is possible that had the sample size been bigger, this participant would have been an outlier. Therefore it is likely that this participant is not a good representation of the rest of the population. The very high pitch also likely contributes to a worse performance of the automatic speech recognition system.

Second of all is the fact that while it was tried to simulate conversational speech and child directed speech. It was not actual conversational speech and child directed speech. Had the speech actually been conversational speech it might be possible that the performance of the system would have been worse, because there would be more stumbling across words. Although simulated child directed speech still differs significantly from adult directed speech (Fernald & Simon, 1984) it is also not

completely the same as actual child directed speech (Fernald & Simon, 1984). Had the speech been actual child directed speech the results could have been different than the results found here.

Third of all, in the evaluation of this system there was no distinction made between words that were very similar to the actual words and words that were not similar to the actual words. For example if the actually said word is “gegaan” and the system transcripted this as “gaan” than this is not less wrong then when the system transcripted this word as “banaan”, even though the word “gaan” is closer in meaning to “gegaan” than the word “banaan”.

Finally, the used materials had their limitations. It was unclear what the ratio of male and female speech was that the used automatic speech recognition system was trained on. So it is concluded that male speech is recognized better because it is likely that that is what the system was trained more on, but it can not be concluded with absolute certainty. Furthermore the audio was recorded using a speech recorder app on a phone, because the experiment was conducted at the home of the experimenter. Thus a recording app on a phone was the only available equipment to record the audio. If better recording equipment had been used, the quality of the audio would have been better. Currently the quality of the audio was not great and it might have interfered with the ability of the system to recognize the speech.

5.3. Further research

The limitations named in the previous section mean that there is room for improvement. So if this subject was researched further there are a number of ways to improve. The first way to improve this current research is to use a bigger sample size. This would make it possible to conduct a significance test to know if the differences between the conditions are actually significant or not. A bigger sample size would also make the results generalizable, which is not the case now. Another way to improve current research is to use multiple automatic speech recognition systems. This would allow you to be able to derive a more general conclusion about the performance of automatic speech recognition systems. Currently this is not possible because only one automatic speech recognition system was evaluated.

Furthermore as mentioned in the previous section, both conversational speech and child directed speech were simulated in this thesis. In further research actual conversational speech and actual child directed speech could be used to research the performance of an automatic speech recognition system. It could also be used to research if the performance significantly differs from cases where it is simulated. Also as mentioned in the previous section, no distinction was made between similar and

dissimilar words. In further research a distinction should be made between such cases, perhaps this could be done by giving a certain score to the transcripted words.

It could also be interesting to train an automatic speech recognition system equally on male and female voices and then research how it performs on adult directed speech and child directed speech. Would there still be a difference in performance between male and female voices, would you still conclude that male child directed speech is better recognized than female child directed speech? Or would the performance be equal? Finally, as mentioned in section 5.1, more research has to be conducted on automatic speech recognition systems and speech register, because currently it is unclear whether training a system on adult directed speech positively or negatively impacts performance on child directed speech.

6. Conclusion

The main research question is “What is the effect of speaker gender and speech register on the performance of an automatic speech recognition system?”. To answer this question, first an answer must be given to the two sub-questions.

“Is the automatic speech recognition system better at recognizing a male voice or a female voice?”. When looking at the results you can conclude that the both adult directed speech and child directed speech are best recognized when used by a male voice. The evaluation values of accuracy, precision and recall are all higher for the male voice than for the female voice. So the system is better at recognizing the male voice. This conclusion is in line with the second hypothesis made in the introduction, so it can be assumed that the system was trained more on male voices than female voices. However it is important to note that in the adult directed speech condition the difference in accuracy, precision and recall between the male and female participant is very small. This difference is a lot bigger in the child directed speech condition.

“Is the automatic speech recognition system better at recognizing child directed speech or is it better at recognizing adult directed speech?”. When looking at the results it is clear that the automatic speech recognition system is best at recognizing child directed speech. The evaluation values of accuracy, precision and recall are all higher for child directed speech than for adult directed speech. So the system is better at recognizing child directed speech. This conclusion is in line with the second hypothesis made in regard to this sub-question in the introduction. So the system is better at recognizing child directed speech because it is slower and more exaggerated.

The results show that the system is best at recognizing male speech and child directed speech. Accuracy, precision and recall are the highest for male child directed speech. Thus the speaker being male and using child directed speech has the most positive effect on the performance of the system, because these two factors are the most positive. This is interesting because it means that a higher pitch has a positive effect, but only within a certain range. Since the highest pitch of the samples, a female speaker using child directed speech, does not have the most positive effect on performance. High pitch only partly positively influences the performance. Another factor that most likely contributed to the good performance of the system on male child directed speech is the duration of the speech, because male child directed speech had the longest duration out of all the samples.

So the following conclusion can be drawn to answer the question *“What is the effect of speaker gender and speech register on the performance of an automatic speech recognition system?”.*

Speaker gender can either have a positive effect on the performance of an automatic speech recognition system, or it can have a more negative effect. If the system is not trained equally on both genders, the gender that the system was trained on more will have a positive effect on the performance and the gender that the system was trained on less will have a more negative effect on the performance. In this case a male speaking voice has a positive effect on the performance of the system, while a female speaking voice has a more negative effect.

Speech register can also either have a positive or a negative effect on the performance of an automatic speech recognition system. If the used speech register is adult directed speech, it negatively influences the performance of the system. If the used speech register is child directed speech, it positively influences the performance of the system.

7. References

Biemans, M. (1998). The effect of biological gender (sex) and social gender (gender identity) on three pitch measures. *Linguistics in the Netherlands*, *15*(1), 41-52.

Broesch, T., & Bryant, G. A. (2018). Fathers' Infant‐Directed Speech in a Small‐Scale Society. *Child development*, *89*(2), e29-e41.

Cooper, R. P., & Aslin, R. N. (1990). Preference for infant‐directed speech in the first month after birth. *Child development*, *61*(5), 1584-1595.

Developers.google. (z.d.). Classification: Accuracy. Retrieved from https://developers.google.com/machine-learning/crash-course/classification/accuracy

Fernald, A., & Mazzie, C. (1991). Prosody and focus in speech to infants and adults. *Developmental psychology*, *27*(2), 209-221.

Fernald, A., & Simon, T. (1984). Expanded intonation contours in mothers' speech to newborns. *Developmental psychology*, *20*(1), 104-113.

Fernald, A., Taeschner, T., Dunn, J., Papousek, M., de Boysson-Bardies, B., & Fukui, I. (1989). A cross-language study of prosodic modifications in mothers' and fathers' speech to preverbal infants. *Journal of child language*, *16*(3), 477-501.

Grieser, D.L., Kuhl, P.K. (1988). Maternal speech to infants in a tonal language: Support for universal prosodic features of Motherese. *Developmental Psychology, 24*(1), 14–20.

Kirchhoff, K., & Schimmel, S. (2005). Statistical properties of infant-directed versus adult-directed speech: Insights from speech recognition. *The Journal of the Acoustical Society of America*, *117*(4), 2238-2246.

Kuhl, P. K., Andruski, J. E., Chistovich, I. A., Chistovich, L. A., Kozhevnikova, E. V., Ryskina, V. L., Stolyarova, E. I., Sundberg, U., and Lacerda, F. (1997). Cross-language analysis of phonetic units in language addressed to infants. *Science, 277*(5326), 684–686

Lands.let.ru (z.d.). Corpusopbouw. Retrieved from http://lands.let.ru.nl/cgn/doc_Dutch/topics/design/design.htm#intro

Lingua. (z.d.). Ruzie in de supermarkt. Retrieved from https://lingua.com/nl/nederlands/lezen/ruzie/

Ludusan, B., Cristia, A., Martin, A., Mazuka, R., & Dupoux, E. (2016). Learnability of prosodic boundaries: Is infant-directed speech easier? *The Journal of the Acoustical Society of America*, *140*(2), 1239-1250.

Maye, J., Werker, J. F., & Gerken, L. (2002). Infant sensitivity to distributional information can affect phonetic discrimination. *Cognition*, *82*(3), B101-B111.

Mendoza, E., Valencia, N., Muñoz, J., & Trujillo, H. (1996). Differences in voice quality between men and women: Use of the long-term average spectrum (LTAS). *Journal of voice*, *10*(1), 59-66.

Openspraaktechnologie. (z.d.). Download. Retrieved from https://www.scribbr.nl/apa-stijl/internetbron-zonder-auteur-datum-titel/

Perault, R., Shoham, Y., Brynjolfsson, E., Clark, J., Etchemendy, J., Grosz, B., Lyons, T., Manyika, J., Niebles, J.C., & Mishra, S. (2019). The AI Index 2019 Annual Report. *AI Index Steering Committee, Human-Centered AI Institute, Stanford University, Stanford, CA*.

Ping Shung, K. (2018, March 15). Accuracy, Precision, Recall or F1? Retrieved from https://towardsdatascience.com/accuracy-precision-recall-or-f1-331fb37c5cb9

Rev. (n.d.). What is WER? What does word error rate mean? Retrieved from https://www.rev.com/blog/resources/what-is-wer-what-does-word-error-rate-mean

Rodger, J. A., & Pendharkar, P. C. (2004). A field study of the impact of gender and user's technical experience on the performance of voice-activated medical tracking application. *International Journal of Human-Computer Studies*, *60*(5-6), 529-544.

Rose, P. (1991). How effective are long term mean and standard deviation as normalisation parameters for tonal fundamental frequency? *Speech Communication*, *10*(3), 229-247.

Sha, F., & Saul, L. K. (2007). Large margin hidden Markov models for automatic speech recognition. In *Advances in neural information processing systems* (pp. 1249-1256).

Simpson, A. P. (2009). Phonetic differences between male and female speech. *Language and linguistics compass*, *3*(2), 621-640.

Tatman, R. (2017, April). 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).

Thiessen, E. D., Hill, E. A., & Saffran, J. R. (2005). Infant-directed speech facilitates word segmentation. *Infancy*, *7*(1), 53-71.

Titze, I. R. 1989. Physiologic and acoustic differences between male and female voices. *The Journal of the Acoustical Society of America 85*(4), 1699–1707.

Werker, J. F., Gilbert, J. H., Humphrey, K., & Tees, R. C. (1981). Developmental aspects of cross-language speech perception. *Child development*, 349-355.

Werker, J. F., Pons, F., Dietrich, C., Kajikawa, S., Fais, L., & Amano, S. (2007). Infant-directed speech supports phonetic category learning in English and Japanese. *Cognition*, *103*(1), 147-162.

Xiong, W., Wu, L., Alleva, F., Droppo, J., Huang, X., & Stolcke, A. (2018). The Microsoft 2017 conversational speech recognition system. In *2018 IEEE international conference on acoustics, speech and signal processing (ICASSP)* (pp. 5934-5938). IEEE.

8. Appendix

8.1. Original text

Ruzie in de supermarkt (Lingua, z.d.)

Gisteren ging ik met Marcel naar de supermarkt. We hadden al een tijd ruzie. Toch moesten er boodschappen worden gedaan, dus besloten we maar om gewoon wat inkopen te gaan doen. We kregen al meer ruzie toen we besloten naar welke supermarkt te gaan. Marcel wilde namelijk naar de supermarkt waar wasmiddel in de aanbieding was en ik wilde naar de supermarkt die het dichtst bij was.

Uiteindelijk stemde ik in met Marcel. We hadden weinig eten in huis dus moesten we best veel kopen. We hadden vooral fruit en verzorgingsproducten nodig.

We kregen alweer ruzie toen we de supermarkt ingingen. Aan het begin van de supermarkt is al het groente en fruit. We moesten besluiten wat we gingen eten die avond, maar we konden het er niet over eens worden. Uiteindelijk hebben we dan ook geen avondeten gekocht.

Nadat we afgerekend hadden waren we nog steeds boos op elkaar. Toch hadden we best veel honger. We zijn toen maar naar de snackbar gegaan voor wat friet.

8.2. Text transcriptions

8.2.1. Male adult directed speech

Gisteren ging met marcel naar de supermarkt tijd ruzie toch moest boodschappen worden gedaan. Dus besloten we maan gewandeld inkoopt gaan doen we kregen ruzie toen we besloten naar welke supermarkt gaan. Marcel wilde namelijk naar de supermarkt waar wasmiddel in de aanbieding was en ik wilde naar de supermarkt die dichtbij was

stemde in met marcel we hadden weinig eten in huis dus toen moesten het best wel veel kopen hadden vooral fruit en verzorgingsproducten nodig.

We kregen alweer ruzie toen we de supermarkt in gingen aan het begin van de supermarkt is al wat groente en fruit en moesten besluiten wat we gingen eten die avond maar we konden er niet over eens worden uiteindelijk hebben we dan ook geen avondeten gekocht.

Nadat we afgerekend hadden waren we nog steeds boos op elkaar toch al wat het best hebben honger we zijn toen maar naar de snackbar gaan bijvoorbeeld friet.

8.2.2. Male child directed speech

Gisteren ging ik met marcel maar de supermarkt badan tijd ruzie toch moest de boodschappen worden gedaan dus besloten we maar gewoon wat inkopen te gaan doen. We kregen al meer ruzie toen besloten naar welke supermarkt gaan marcel wilde namelijk naar de supermarkt waar wasmiddel in de aanbieding was en ik wilde naar de supermarkt die dichtst bij was

uiteindelijk stemde ik in met marcel. We hadden weinig eten in huis dus we moesten best veel kopen we hadden vooral fruit en verzorgingsproducten nodig.

We kregen al weer ruzie toen we de supermarkt in gingen aan het begin van de supermarkt is al groente en fruit en moesten besluiten wat we gingen eten die avond maar we konden daar niet over eens worden uit eindelijk hebben we dan ook geen avondeten gekocht.

Nadat we afgerekend hadden waren we nog steeds boos op elkaar toch hadden we best veel honger we zijn toen maar naar de snackbar gegaan voor friet.

8.2.3. Female adult directed speech

Gisteren ging ik bij marcel naar de supermarkt want altijd ruzie toch moesten er boodschappen worden gedaan dus besloot om maar gewoon wat inkopen gaan doen we kregen om meer ruzie toen besloten naar welke supermarkt te gaan. Maar ze hadden naar de supermarkt waar het was midden in de aanbieding was en ik wilde naar de supermarkt die dichtbij was

uiteindelijk stemde ik in met marshall hadden weinig eten in huis dus we moesten het best verkopen hadden we vooral fruit en verzorgingsproducten nodig

we kregen alweer ruzie toen we de supermarkt in gingen. Aan het begin van de supermarkt groente en fruit moesten besluiten wat we gingen eten die avond komt er niet over eens worden uiteindelijk heb ik dan ook geen avondeten gekocht

nadat we afgerekend hadden waren we nog steeds boos op elkaar toch hadden we best veel honger we zijn toen maar naar de snackbar gegaan voordat friet.

8.2.4. Female child directed speech

Gisteren ging ik met marcel naar de supermarkt we hadden altijd ruzie toch moest zou ik boodschappen worden gedaan dus besloten om maar gewoon wat inkopen gaan doen we kregen wel meer ruzie toen besloten naar welke supermarkt te gaan. Marcha wilde namelijk naar de supermarkt waar het wasmiddel in de aanbieding was en ik wilde naar de supermarkt die dichtbij was

uiteindelijk stemde in met marcel. We hadden weinig eten in huis dus moest ze best veel kopen. Wonen vooral fruit en verzorgingsproducten nodig.

We kregen al veel ruzie toen we de supermarkt gingen aan het begin van de supermarkt is altijd groente en fruit moesten besluiten wat we gingen eten die avond maar kon ze niet over eens worden uiteindelijk hebben we dan ook geen eten gekocht

nadat we afgerekend hadden waren we nog steeds boos op elkaar. Toch hadden we best veel honger zijn doen maar naar de snackbar gegaan voor wat friet.

8.3. Text evaluations

8.3.1. Male adult directed speech

Input speaker	Output system	True positives (cumulative)	False negatives (cumulative)	False positives (cumulative)
Gisteren ging	Gisteren ging	2	-	-
ik	-	2	1	-
met Marcel naar de supermarkt	met Marcel naar de supermarkt	7	1	-
We hadden al een	-	7	5	-
tijd ruzie toch	tijd ruzie toch	10	5	-
moesten	moest	10	6	1
er	-	10	7	1
boodschappen worden gedaan, dus besloten we	boodschappen worden gedaan, dus besloten we	16	7	1
maar	maan	16	8	2
om gewoon wat	-	16	11	2
-	gewandeld	16	11	3
inkopen	inkoopt	16	12	4
te	-	16	13	4
gaan doen. We kregen	gaan doen. We kregen	20	13	4
al meer	-	20	15	4
ruzie toen we besloten naar welke supermarkt	ruzie toen we besloten naar welke supermarkt	27	15	4
te	-	27	16	4
gaan. Marcel wilde namelijk naar de supermarkt waar wasmiddel in de aanbieding was en ik wilde naar de supermarkt die	gaan. Marcel wilde namelijk naar de supermarkt waar wasmiddel in de aanbieding was en ik wilde naar de supermarkt die	47	16	4

het	-	47	17	4
dichtst bij	dichtbij	47	19	5
was	was	48	19	5
Uiteindelijk	-	48	20	5
stemde	stemde	49	20	5
ik	-	49	21	5
in met Marcel. We hadden weinig eten in huis dus	in met Marcel. We hadden weinig eten in huis dus	59	21	5
-	toen	59	21	6
moesten	moesten	60	21	6
we	het	60	22	7
best wel veel kopen	best wel veel kopen	64	22	7
we	-	64	23	7
hadden vooral fruit en verzorgingsproducten nodig. We kregen alweer ruzie toen we de supermarkt ingingen. Aan het begin van de supermarkt is al	hadden vooral fruit en verzorgingsproducten nodig. We kregen alweer ruzie toen we de supermarkt ingingen. Aan het begin van de supermarkt is al	87	23	7
het	wat	87	24	8
groente en fruit	groente en fruit	90	24	8
We	en	90	25	9
moesten besluiten wat we gingen eten die avond, maar we konden	moesten besluiten wat we gingen eten die avond, maar we konden	101	25	9
het	-	101	26	9
er niet over eens worden. Uiteindelijk hebben we dan ook geen avondeten gekocht. Nadat we	er niet over eens worden. Uiteindelijk hebben we dan ook geen avondeten gekocht. Nadat we	126	26	9

hadden afgerekend waren we nog steeds boos op elkaar. Toch	hadden afgerekend waren we nog steeds boos op elkaar. Toch			
hadden we	al wat het	126	28	12
best	best	127	28	12
veel	hebben	127	29	13
honger. We zijn toen maar naar de snackbar	honger. We zijn toen maar naar de snackbar	135	29	13
gegaan	gaan	135	30	14
voor wat	bijvoorbeeld	135	32	15
friet	friet	136	32	15

8.3.2. Male child directed speech

Input speaker	**Output system**	**True positives (cumulative)**	**False negatives (cumulative)**	**False positives (cumulative)**
Gisteren ging ik met Marcel	Gisteren ging ik met Marcel	5	-	-
naar	maar	5	1	1
de supermarkt	de supermarkt	7	1	1
We hadden al een	-	7	5	1
-	badan	7	5	2
tijd ruzie. Toch	tijd ruzie. Toch	10	5	2
moesten	moest	10	6	3
er	de	10	7	4
boodschappen worden gedaan, dus besloten we maar	boodschappen worden gedaan, dus besloten we maar	17	7	4
om	-	17	8	4
gewoon wat inkopen te gaan doen. We	gewoon wat inkopen te gaan doen. We	29	8	4

kregen al meer ruzie toen	kregen al meer ruzie toen			
we	-	29	9	4
besloten naar welke supermarkt	besloten naar welke supermarkt	33	9	4
te	-	33	10	4
gaan. Marcel wilde namelijk naar de supermarkt waar wasmiddel in de aanbieding was en ik wilde naar de supermarkt die	gaan. Marcel wilde namelijk naar de supermarkt waar wasmiddel in de aanbieding was en ik wilde naar de supermarkt die	53	10	4
het	-	53	11	4
dichtst bij was. Uiteindelijk stemde ik in met Marcel. We hadden weinig eten in huis dus we moesten ~~we~~ best veel kopen. We hadden vooral fruit en verzorgingsproducten nodig. We kregen alweer ruzie toen we de supermarkt ingingen. Aan het begin van de supermarkt is al	dichtst bij was. Uiteindelijk stemde ik in met Marcel. We hadden weinig eten in huis dus we moesten best veel kopen. We hadden vooral fruit en verzorgingsproducten nodig. We kregen alweer ruzie toen we de supermarkt ingingen. Aan het begin van de supermarkt is al	98	11	4
het	-	98	12	4
groente en fruit.	groente en fruit.	101	12	4
We	en	101	13	5
moesten besluiten wat we gingen eten die avond, maar we konden	moesten besluiten wat we gingen eten die avond, maar we konden	112	13	5

het er	daar	112	15	6
niet over eens worden. Uiteindelijk hebben we dan ook geen avondeten gekocht. Nadat we afgerekend hadden waren we nog steeds boos op elkaar. Toch hadden we best veel honger. We zijn toen maar naar de snackbar gegaan voor	niet over eens worden. Uiteindelijk hebben we dan ook geen avondeten gekocht. Nadat we afgerekend hadden waren we nog steeds boos op elkaar. Toch hadden we best veel honger. We zijn toen maar naar de snackbar gegaan voor	150	15	6
wat	-	150	16	6
friet	friet	151	16	6

8.3.3. Female adult directed speech

Input speaker	**Output system**	**True positives (cumulative)**	**False negatives (cumulative)**	**False positives (cumulative)**
Gisteren ging ik	Gisteren ging ik	3	-	-
met	bij	3	1	1
Marcel naar de supermarkt	Marcel naar de supermarkt	7	1	1
We hadden al een tijd	-	7	6	1
-	want altijd	7	6	3
ruzie. Toch moesten er boodschappen worden gedaan, dus	ruzie. Toch moesten er boodschappen worden gedaan, dus	15	6	3
besloten	besloot	15	7	4
we	-	15	8	4
om maar ~~om~~ gewoon wat inkopen	om maar gewoon wat inkopen	20	8	4
te	-	20	9	4

gaan doen. We kregen	gaan doen. We kregen	24	9	4
al	om	24	10	5
meer ruzie toen	meer ruzie toen	27	10	5
we	-	27	11	5
besloten naar welke supermarkt te gaan.	besloten naar welke supermarkt te gaan.	33	11	5
Marcel wilde namelijk	-	33	14	5
-	Maar ze hadden	33	14	8
naar de supermarkt waar	naar de supermarkt waar	37	14	8
wasmiddel	-	37	15	8
-	het was midden	37	15	11
in de aanbieding was en ik wilde naar de supermarkt die	in de aanbieding was en ik wilde naar de supermarkt die	48	15	11
het	-	48	16	11
dichts bij	dichtbij	48	18	12
was. Uiteindelijk stemde ik in met	was. Uiteindelijk stemde ik in met	54	18	12
Marcel	Marshall	54	19	13
We	-	54	20	13
hadden weinig eten in huis dus we moesten ~~we~~	hadden weinig eten in huis dus we moesten	62	20	13
-	het	62	20	14
best	best	63	20	14
veel kopen	verkopen	63	22	15
We	-	63	23	15
hadden	hadden	64	23	15

-	we	64	23	16
vooral fruit en verzorgingsproducten nodig. We kregen alweer ruzie toen we de supermarkt ingingen. Aan het begin van de supermarkt	vooral fruit en verzorgingsproducten nodig. We kregen alweer ruzie toen we de supermarkt ingingen. Aan het begin van de supermarkt	84	23	16
is al het	-	84	26	16
groente en fruit	groente en fruit	87	26	16
We	-	87	27	16
moesten besluiten wat we gingen eten die avond,	moesten besluiten wat we gingen eten die avond,	95	27	16
maar we konden het	-	95	31	16
-	komt	95	31	17
er niet over eens worden. Uiteindelijk	er niet over eens worden. Uiteindelijk	101	31	17
hebben we	heb ik	101	33	19
dan ook geen avondeten gekocht. Nadat we afgerekend hadden waren we nog steeds boos op elkaar. Toch hadden we best veel honger. We zijn toen maar naar de snackbar gegaan	dan ook geen avondeten gekocht. Nadat we afgerekend hadden waren we nog steeds boos op elkaar. Toch hadden we best veel honger. We zijn toen maar naar de snackbar gegaan	131	33	19
voor wat	voordat	131	35	20
friet	friet	132	35	20

8.3.4. Female child directed speech

Input speaker	Output system	True positives	False negatives	False positives

		(cumulative)	(cumulative)	(cumulative)
Gisteren ging ik met Marcel naar de supermarkt. We hadden	Gisteren ging ik met Marcel naar de supermarkt. We hadden	10	-	-
al een tijd	altijd	10	3	1
ruzie. Toch	ruzie. Toch	12	3	1
moesten	moest	12	4	2
er	-	12	5	2
-	zou ik	12	5	4
boodschappen worden gedaan, dus besloten	boodschappen worden gedaan, dus besloten	17	5	4
we	-	17	6	4
om maar ~~om~~ gewoon wat inkopen	om maar gewoon wat inkopen	22	6	4
te	-	22	7	4
gaan doen. We kregen	gaan doen. We kregen	26	7	4
al	wel	26	8	5
meer ruzie toen	meer ruzie toen	29	8	5
we	-	29	9	5
besloten naar welke supermarkt te gaan	besloten naar welke supermarkt te gaan	35	9	5
Marcel	Marcha	35	10	6
wilde namelijk naar de supermarkt waar het wasmiddel in de aanbieding was en ik wilde naar de supermarkt die	wilde namelijk naar de supermarkt waar het wasmiddel in de aanbieding was en ik wilde naar de supermarkt die	54	10	6
het	-	54	11	6
dichtst bij	dichtbij	54	13	7

was. Uiteindelijk stemde	was. Uiteindelijk stemde	57	13	7
ik	-	57	14	7
in met Marcel. We hadden weinig eten in huis dus	in met Marcel. We hadden weinig eten in huis dus	67	14	7
we	-	67	15	7
moesten	moest	67	16	8
-	ze	67	16	9
best veel kopen	best veel kopen	70	16	9
We hadden	wonen	70	18	10
vooral fruit en verzorgingsproducten nodig. We kregen	vooral fruit en verzorgingsproducten nodig. We kregen	77	18	10
alweer	al veel	77	19	12
ruzie toen we de supermarkt	ruzie toen we de supermarkt	82	19	12
ingingen	gingen	82	20	13
Aan het begin van de supermarkt is	Aan het begin van de supermarkt is	89	20	13
al het	altijd	89	22	14
groente en fruit.	groente en fruit.	92	22	14
we	-	92	23	14
moesten besluiten wat we gingen eten die avond, maar	moesten besluiten wat we gingen eten die avond, maar	101	23	14
we	-	101	24	14
konden het er	kon ze	101	27	16

niet over eens worden. Uiteindelijk hebben we dan ook geen	niet over eens worden. Uiteindelijk hebben we dan ook geen	111	27	16
avondeten	eten	111	28	17
gekocht. Nadat we afgerekend hadden waren we nog steeds boos op elkaar. Toch hadden we best veel honger.	gekocht. Nadat we afgerekend hadden waren we nog steeds boos op elkaar. Toch hadden we best veel honger.	129	28	17
We	-	129	29	17
zijn	zijn	130	29	17
toen	doen	130	30	18
maar naar de snackbar gegaan voor wat friet	maar naar de snackbar gegaan voor wat friet	138	30	18