Prosodic modification of infant-directed speech for improved automatic speech recognition

Koen Knape

Supervisor: Anika van der Klis

02-06-2021

BA Linguistics

Utrecht University

Contents

Abstract

The current performance of automatic speech recognition (ASR) for infant-directed speech (IDS) leaves a lot to be desired. If this performance could be increased to more reliable levels, it would enable researchers to use automatic speech recognisers for IDS research, which would drastically decrease the time needed for research on IDS. This thesis focuses on possible improvements for this problem, utilizing a front-end approach. IDS-fragments were first prosodically modified for average pitch and speaking rate, and were subsequently fed to an ASR-system. The resulting automatic annotations were then evaluated for their precision, recall and F-score. However, no improvements of the ASR-system's performance were found after prosodic modification. Although these results did not show improved performance of the recognition system, this study does provide some interesting insights for future research regarding ASR for IDS.

1. Introduction

Infant-directed speech (IDS), also known as *motherese* or *babytalk,* is an important part of the linguistic input of younger children. Research on the exact linguistic properties of this speech register ranges primarily between levels of phonology, syntax, pragmatics and lexical use (Fernald & Simon, 1984; Fernald et al., 1989; Cooper & Aslin, 1990; Soderstrom, 2007; Cristia & Seidl, 2014). While IDS shows interesting variation in all of these subfields, this thesis specifically focuses on the prosodic properties of Dutch IDS in relation to the performance of the automatic speech recognition (ASR) system Kaldi-NL. The typical prosodic features of IDS include a higher mean pitch, a wider pitch range, and a slower speaking rate compared to adult-directed speech (ADS). This divergence may be one of the reasons why ASR-performance for IDS is less accurate than ASR of ADS (see Kirchhoff & Schimmel, 2005; Van der Klis et al., 2020). The ultimate goal of the study was to improve this performance by adjusting the pitch and speaking rate of fragments of recorded IDS before feeding them to the ASR-system. If such improvements could be found, ASR would be a step closer to reliable application for IDS, which could drastically speed up the time that is currently needed for IDS research.

1.1. Structure of this thesis

This thesis is organized as follows. Chapter 2 first discusses the theoretical background and then posits the research questions and hypotheses. In chapter 3, the explanation of the methodology is provided; the data, materials and approach of the methodology are introduced, as well as the means for evaluating the performance of the ASR-system. Then, in chapter 4, the results of the evaluation are presented. Chapter 5 discusses these findings and relates them to the existing literature, and uses them to answer the research questions. Finally, chapter 6 concludes this thesis and offers inspiration for future research.

2. Theoretical background

2.1. Properties of infant-directed speech

One of the most prominent features of IDS might be the exaggerated prosodic patterns in which parents talk to their infant. These patterns include exaggerated levels of pitch (Fernald et al., 1989) and slower speaking rates (Fernald & Simon, 1984; Potamianos et al., 1997). Pitch shows a higher average fundamental frequency (F0) as well as a higher degree of variability (the difference between the minimum and maximum F0 values) compared to ADS (Fernald & Simon, 1984; Fernald et al., 1989; Cooper & Aslin, 1990).

Fernald and Simon (1984) examine in their research these prosodic properties of IDS. They conducted an experiment with 24 German mothers, in two different conditions: in the one condition they spoke to their 3- to 5-day-old baby and in the other condition they spoke to an adult interviewer. In the IDS condition, they found that the average F0 increased by 27% to 257 Hz, as compared to ADS, where the F0 was 203 Hz. The speaking rate showed a clear difference too and slowed down with 28%: 4.2 syllables per second in IDS versus 5.8 syllables per second in ADS. In a similar study, Fernald et al. (1989) studied IDS in a broader context, by comparing the effects of speaker gender and the language spoken on IDS. Mothers as well as fathers spoke several different languages and language variants. The children were all aged between 10 and 14 months, and they were spoken to by their parents. These results show a consistent increase in pitch levels in IDS of the parents as well, although mothers' average pitch was higher than that of the father: German mothers increased their pitch 16% in IDS, compared to a 7% increase of the fathers. This is a more moderate increase than reported in Fernald and Simon (1984), which might be explained by the age-related effects of the children.

Similar prosodic exaggerations of IDS can be found in many languages, such as shown in Fernald et al.'s (1989) study, although differences in specific IDS characteristics across languages exist. Not all languages adapt the properties of speech when speaking to children to the same extent, or even adapt the same properties at all. This language specificity in IDS includes differences in pitch. For example, tonal languages like Mandarin Chinese use pitch to discriminate lexical meaning. This might conflict with IDS' modifications of pitch (Soderstrom, 2007). On the other hand, there are languages like English and Dutch which use lexical stress instead, leaving a higher degree of freedom for pitch-modification. In American English, these modifications are most notable (Fernald et al., 1989): speakers tend to use the highest mean F0 and the highest pitch variability. Furthermore, the vowel space in IDS is increased, meaning that vowels are generally more phonetically distinct (Kirchhoff and Schimmel, 2005; Han, 2019). However, such "hyperarticulation" is not consistent across *all* speech sounds. For example, Cristia and Seidl (2014) argue that perhaps only the *point-vowels* are hyperarticulated, and that other vowels are left unenhanced. On the other hand, Dutch IDS shows tendencies in the opposite direction. In Benders (2013), the author shows results that oppose Cristia and Seidl's (2014) hypothesis of hyperarticulated point-vowels. She studied speech sound contrasts and whether these are language-universal, and showed results that indicated a *decreased* vowel space between the point-vowels in Dutch IDS, indicating *hypo*articulation instead of *hyper*articulation. These studies' findings illustrate the cross-linguistic differences in the prosodic and acoustic realization of IDS.

The age of children that speech is directed at plays a role in the realization of the prosodic properties of IDS. Children at younger ages have rapid developing language competencies, which may influence how IDS is realized. The general trend is that the older infants get, the more the IDS begins to resemble usual ADS (Han, 2019). This trend can be seen in the above comparison of the average pitch of IDS in the studies of Fernald and Simon (1984) and Fernald et al. (1989), where pitch was increased more for the younger children. However, such trends are not always linear. This is illustrated by Kitamura et al. (2002), who showed in their research an increase of pitch in IDS at 6 months, a decrease at 9 months, and then again an increase at 12 months. The same trend shows in Benders (2013), where pitch-levels for the 15-month-olds were higher than those for the 11-month-olds. These results illustrate the effects of age on the pitch in IDS, and how the effect exhibits different patterns at different ages. While age-related effects on pitch are documented extensively, less attention has been paid to age-related effects on speaking rate (Han, 2019).

IDS shows differences compared to ADS when it comes to prosodic properties. Pitch reaches higher average values, but apart from that, it also shows greater variability (Cristia & Seidl, 2014; Miyazawa et al., 2017). Studies show varying results when measuring F1 and F2 values of vowels. The general trend is that in IDS the values of these first two formants are generally more distinct between categories (Cristia & Seidl, 2014; Kirchhoff & Schimmel, 2005). But not only the difference *between* vowels is generally increased; the difference *within* vowels is increased as well. This means that in IDS, vowels are articulated with a higher degree of variability than those same vowels in ADS. Consequently, categories in IDS can overlap more than in ADS. This causes it to be harder to make a distinction between vowels, thus making them harder to learn, which is especially true for categories that are close together on the formant space. So, variability in IDS can both be an asset as well as a hindrance when learning a language.

2.1. Automatic speech recognition of infant-directed speech

Automatic speech recognition systems have the goal of mapping speech sound input to an orthographic representation of the utterance. Such a system is generally based on two main models. One is a statistical language model that works with probabilities of certain word sequences, and that is trained with corpora of textual sequences. The other one is the acoustic model, which looks at how likely it is that certain textual sequences are realized in a particular acoustic way. This model is trained on collections of acoustic data with accompanying transcriptions. Together, these two models shape the system. This is then used to determine the most probable transcription that would fit a given speech sound input.

The data these models are trained on has its effect on what data the system recognizes best. In Kirchhoff and Schimmel (2005), the authors make a distinction between two different kinds of acoustic models that can underly ASR-systems, namely one that is trained on IDS, and one that is trained on ADS. As described in the sections before, these are phonetically very different speech registers. In their study, Kirchhoff and Schimmel (2005) compare the performance of speech recognition for matched and mismatched conditions for the IDS and ADS register. In matched conditions, an ASR-system with models trained on one register was applied to that same register. In mismatched conditions, the system's model was trained on one register and applied to the other. Their resulting accuracy scores are displayed in table 1. The results show the accuracy scores of the two different models when applied to IDS: the model trained on ADS scored 91.8%, while the model trained on IDS scored 94.7%.

As can be seen from these results, models trained on one speech register recognize that same register best, while its performance deteriorates when applied to speech with different acoustical properties. Thus, prosodic divergences of IDS compared to ADS can cause problems when it needs to be recognised by ASR-systems. Although this study shows that training the acoustic model of ASR on IDS data can improve its results when recognising IDS, it still performs worse than ASR-systems trained on ADS applied to ADS. The discrepancies between IDS and ADS in these matched performances are not very explicitly discussed in Kirchhoff and Schimmel's study (2005). Nevertheless, one could argue that IDS' variability could be correlated with this discrepancy. If a recognition system is trained on datasets with high variability, e.g. IDS, this means that there is relatively more overlap between speech sound classes than when trained on less variable datasets. This would result in poorer separability of these speech sounds.

		Test	
		ADS	**IDS**
Train	**ADS**	97.4%	91.8%
	IDS	92.9%	94.7%

Table 1. *Accuracy scores for IDS- and ADS-trained ASR-systems under matched and mismatched conditions (Kirchhoff & Schimmel, 2005).*

The findings of Van der Klis et al.'s study (2020) further indicate the worse performances of ASR for IDS. In their research, the authors assessed the performance of ASR for IDS at 18 months and IDS at 24 months, and compared this to the ASR-performance for ADS. The resulting evaluation scores are displayed in table 2. The worse ASR-performance for IDS is clearly illustrated by these results and shows the strongest for the younger age-groups, where recall for IDS is almost 16% lower than for ADS.

	18 months		24 months	
Register	**ADS**	**IDS**	**ADS**	**IDS**
Recall	71.0%	55.4%	60.5%	53.8%
Precision	100%	100%	100%	100%
F-score	83.3%	71.3%	75.4%	70.0%

Table 2. *Evaluation scores for ASR-performance for IDS and ADS across 18- and 24-months (Van der Klis et al., 2020).*

Based on Kirchhoff and Schimmel (2005) and Van der Klis et al. (2020), it is clear that the performance of ASR for IDS leaves a lot of room for improvement.

While there is no solution for improving the automatic recognition of IDS yet, helpful direction to look in is at automatic recognition of children's speech (CS). The prosodic characteristics of IDS overlap strongly with those of CS: CS shows a slower speaking rate and higher pitch levels too (Gustafson & Sjölander, 2002; Stemmer et al., 2003; Ghai & Sinha, 2015). Taking this similarity into account means that for improving ASR for IDS, it is also useful to look at research concerning ASR and CS.

Booth et al. (2020) show that training an ASR-model on mixed data sets (containing adult speech as well as CS) improves its performance with a decrease of word error rate (WER) of around 30%. However, this approach is just one of the possible ways to improve ASR-performance. Another way to do so is shown by Stemmer et al. (2003), who adjust the prosodic properties of the speech recording itself before feeding it to a recognizer. The maximum reduction of WER after adjustment of pitch was 25.6%, with the use of non-linear vocal-tract length normalization (VTLN). By scaling down the speaking rate with a factor between 0.84 and 0.88, a maximum WER reduction of even 30.9% was achieved.

Another technique that can be used for acoustic adjustments is with help of a pitch normalization algorithm, such as the Pitch Synchronous Overlap and Add-algorithm, or PSOLA-algorithm (used in Gustafson & Sjölander, 2002 and Stemmer et al., 2003). Age-specific pitch adjustments of ASR-models can also be applied for better results (Hagen et al., 2007), since IDS shows great variation due to age-related effects. For adjustment of speaking rate, a method such as the PSOLA-algorithm is suited as well, as can be seen in the positive results in Stemmer et al.'s research (2003).

Roughly two ways for improving recognition for IDS can be distinguished in the discussed literature. The first one is the back-end approach, where the models that underlie ASR are adapted, e.g., by training them on different data or by adjusting the ASR-models accordingly. The other way is the front-end approach, in which the speech is adapted before being recognized by ASR. This type of modification is a relatively straightforward approach and yields desirable results in enhancing ASR's performance on IDS.

2.2. The current study

There is good reason to improve ASR for registers such as IDS. This has to do with data analysis in research concerning IDS. Such data is in the form of recorded speech-fragments which often comprise of relative long spans of time. Before analysis can be done on these data, the fragments need to be annotated and target segments have to be extracted first. This process not only takes considerable amounts of time, but also requires experience and speech-encoding skills from researchers. By applying automatic methods to perform this process, numerous manual working hours can be cut, which would significantly reduce the total amount of time necessary to conduct such research and to achieve results. Furthermore, researchers would need fewer coding skills and less relevant experience to analyse and annotate speech.

For this application of ASR to IDS to be possible, however, speech recognition first needs to be improved to more reliable levels of recognition. This is what the current study is concerned with. It will assess the performance of ASR on IDS and aims to improve this using a front-end approach, in which speech recordings of IDS will be adapted in their prosodic properties. In combination with the findings in the discussed literature, this leads to the key question of this research: can ASR's performance on IDS be improved by adjustment of its prosodic properties? From this main question, three specific sub-questions follow, which are concerned with the adaptation of the two main prosodic components of speech.

1. Will the lowering of pitch in IDS-fragments lead to better ASR-performance?
2. Will the increase of speaking rate in IDS-fragments lead to better ASR-performance?
3. Will the combination of lowering the pitch and increasing the speaking rate in IDS-fragments lead to better ASR-performance?

The hypothesis is that the respective prosodic adjustments in IDS can lead to a better performance of ASR for IDS. Studies on the improvement of ASR for CS show that such modifications can achieve such improved performances (Gustafson & Sjölander, 2002; Stemmer et al., 2003; Ghai & Sinha, 2015). Since CS shows strong similarities with IDS, these studies' results reinforce the expectation that in the current study, similar modifications of IDS will yield improvements for the performance of ASR as well. Might this hypothesis prove to be true, then this study would provide researchers from all over the world with accessible and quick means to improve automatic recognition of IDS, given that they have access to an ASR-system suited for the relevant language.
On the other hand, it could be the case that such prosodic adjustments to IDS do not lead to improvements for ASR performance. In this case, it might not be the prosodic changes, but the troubles with modelling the higher variability inherent to IDS that lead to negative results. If this null hypothesis would prove to be true, then it might be a good next step for future research to focus on the problem of the variability of IDS.

3. Method

The goal of the current study is to assess the performance of ASR applied to IDS and to find out how this performance could be improved using a front-end approach. This approach consisted of a stepwise adjustment of the prosodic features of IDS-fragments. After these adjustments, the resulting speech-fragments were automatically annotated with the use of an ASR-system. These annotations were evaluated against the manual annotations.

3.1. Data

The speech-fragments used in this study are taken from Mengru Han's dissertation (2019) on the role of prosodic input in word learning. A group of 22 mother-child pairs were studied at two moments in time; once when the children were 18 months old, and again when they were 24 months old. All mothers were native speakers of Dutch who had higher education, and all children were typically developing. This current study only focuses on the data from the 18-month-old group (M = 18.4 months, range = 18.0 – 18.9 months, 11 girls), because at younger ages, IDS typically diverges from ADS the most. Successful modification of these fragments will therefore be most likely to reveal what specific prosodic properties of IDS are most useful to adjust for improved ASR-performance.

In the experiment, the mothers read a picture book in two conditions to record both IDS and ADS. This book was meant to elicit only seven disyllabic target words (table 3), with no further script given. Only these words were studied, because this allowed for a consistent comparison of the same words across different participants, while at the same time retaining a free-speech component in their speech. Every keyword occurred at least once per recording, but the exact number of occurrences varied between fragments, since some of the words were repeated one or more times. The resulting speech-fragments (N = 22) were manually annotated in previous work by Han (2019). These annotations serve as the "golden standard" and are key in the eventual evaluation of ASR-performance. All participants were recorded in a sound-proof in the Utrecht Baby Lab, with the use of a ZOOM H1 recorder (at a sampling rate of 44.1 kHz). The recording sessions lasted about 15-20 minutes each. For further detail on the procedure, see Han (2019).

Target words	English translation
1. appel	"apple"
2. walnoot	"walnut"
3. eland	"moose"
4. bever	"beaver"
5. kasteel	"castle"
6. opa	"grandpa"
7. pompoen	"pumpkin"

Table 3. *Target words in Mengru Han (2019).*

3.2. Determining prosodic modification of speech-fragments

In the current study, the IDS-fragments from Han (2019) were modified in the direction of ADS. This means that their pitch was decreased and that their speaking rate was increased, so that these values became more similar to those typical of ADS. Initially, the IDS-fragments' pitch-decrease was determined to be 23%. This choice was based on the study of Gustafson and Sjölander (2002), who showed that this modification had a positive effect on WER's of their speech recognizer. However, their study was concerned with the modification of CS. Although this register shows strong similarities with IDS, the current study only focuses on the modification of the latter. For this reason, further modifications specifically for IDS-fragments were determined, using the data from Han (2019). First, the average pitch and speaking rate from IDS- and ADS-fragments from Han (2019) were calculated. Next, the proportional differences of these averages were used to determine the eventual prosodic adjustments that were necessary to make the IDS-fragments more similar to ADS.

The averages are represented in table 4. Both pitch and speaking rate were measured at two levels: at word level and utterance level. For speaking rate, the duration of pauses is excluded at the utterance level. In Han (2019), the average prosodic values were further subdivided between two more conditions: familiar and unfamiliar target words. Within each speech fragment, two types of familiar words and five types of unfamiliar words occurred. In this current study, however, this distinction is left out and the weighted average of these two familiarity conditions are used instead. This allows for the speech-fragments to be modified in their entirety, instead of having to adjust words within speech-fragments based on familiarity. These weighted averages are represented in table 4.

		ADS	**IDS**
Average F0	Utterance level	228.82	259.51
	Word level	244.48	261.21
Average speaking rate	Utterance level	5.12	4.66
	Word level	4.93	4.43

Table 4. *Average F0 (Hz) and speaking rate (syllables/second) in ADS versus IDS speech-fragments at utterance and word level as taken from Han (2019).*

Next, the average proportions of ADS and IDS are calculated at the hand of the information in table 4, by dividing average ADS values by average IDS values. The resulting proportions are represented in table 5. Again, there are two proportions per prosodic aspect: one at utterance level and one at word level.

		Proportion
Average F0	Utterance level	0.88
	Word level	0.94
Average speaking rate	Utterance level	1.10
	Word level	1.11

Table 5. *Proportions of ADS and IDS ($\frac{ADS}{IDS}$) for average F0 and speaking rate, on utterance level and word level.*

3.3. Approach

The proportions in table 5 were used to perform a stepwise modification of the speech-fragments, which was done with the use of the open-source audio software Audacity (Audacity Team; version 3.0.2.). This program offers the possibility to modify speaking rate without altering pitch and vice versa. Since in this study ASR-performance is tested solely on recognition of target words, only the proportions for word level were used to modify the IDS-fragments. Thus, the modifications were then made in three steps. First, the modification of pitch based on the findings of Gustafson and Sjölander (2002) for CS was applied. Secondly, the prosodic modifications based on the relevant proportions in table 5 were applied, to adjust the average prosodic measures of IDS to the levels of ADS. The first prosodic modification was the adjustment of the pitch by its corresponding proportion, the second modification was applied to speaking rate and the third modification was applied to both pitch and speaking rate at once. In Audacity, all these modifications could be made by inserting either the relevant percentage or the proportion. This stepwise modification of the fragments resulted in four different sets of speech-fragments, which were then recognized by the ASR-system one by one.

3.4. The automatic speech recognition system: Kaldi-NL

The ASR-system that is used in this study is Kaldi-NL, a state-of-the-art ASR-system that is developed by the Dutch Foundation of Open Speech Technology (https://openspraaktechnologie.org). It consists of software that is built around KALDI (Povey et al., 2011) – which is the base version – and is extensively trained on Dutch language models. These acoustic models of Dutch are developed with help of the Corpus of Spoken Dutch (Nederlandse Taalunie, 2006), which is an exceptionally extensive corpus that contains approximately 1000 hours of Dutch spoken by adults. For the ASR-task in the current study, the parameter of Kaldi-NL was set to "Daily conversations".

3.5. Evaluation of the ASR-system

This ASR-system was applied to the modified versions of the 22 IDS-fragments, and it returned a set of annotations as output. With the help of a script, this set of annotations was compared to the manual annotations, and from this comparison followed the basis for evaluation: the frequency of hits, misses and false positives of the ASR-system. Hits are the correctly identified target words in the automatic annotations; misses are the target words from the speech-fragment that remained unrecognized; false positives are words that are incorrectly recognized as target words. These three categories were used to evaluate the ASR-performance. Three accuracy scores that were calculated for this evaluation are precision, recall and F-score, which are defined as follows:

$$Precision = \frac{hits}{hits + false\ positives}$$

$$Recall = \frac{hits}{hits + misses}$$

$$F = 2 \times \frac{precision \times recall}{precision + recall}$$

Precision measures the accuracy of all of the items that are retrieved by the ASR-system, while recall measures the proportion of target words that are recognized in the data. Finally, the F-score is the harmonic mean of both precision and recall (for further reference on these accuracy scores, see chapter 4.7 of Jurafsky and Martin (2020)).

3.5.1. Morphological variation and normalization

Although all participants used the necessary target words during their recording sessions, they also used several morphological variants of the target words. Diminutive forms, plural forms, and some compounded forms occurred in the speech-fragments. All these types of morphological variants are shown in table 6. These variants accounted for a significant amount of the total amount of target words (roughly 14%). This meant that for the evaluation script to return more accurate scores, the ASR-output had to be adjusted. Accordingly, all morphological variation of target words in the ASR-output was normalized to its keyword form.

Keyword	Morphological variant	Type
Appel (singular)	Appels	Plural
	Appeltje	Diminutive
	Appeltjes	Plural, diminutive
	Appelsap	Compound
	Sinaasappel	Compound
	Sinaasappeltje	Compound, diminutive
	Sinaasappeltjes	Compound, diminutive and plural
	Sinaasappelsap	Compound
	Sinaasappelpartjes	Compound
Walnoot (singular)	Walnoten	Plural
Kasteel (singular)	Kastelen	Plural
Pompoen (singular)	Pompoenen	Plural
	Pompoensoep	Compound

Table 6. *Morphological variation of target words found by the ASR-system.*

4. Results

In total, four kinds of prosodic modifications were applied to the IDS-fragments. These consist of two different modifications of pitch, one modification of speaking rate, and finally one modification of both pitch and speaking rate. Additionally, the target words in the ASR-output were normalized for morphological variation. The evaluation scores of all modifications are represented in table 7. Effects of these modifications on ASR-performance were compared to the baseline, which consists of findings from Van der Klis et al. (2020). These baseline results are displayed in the second row of table 7 and represent the evaluation scores for ASR-performance for unmodified IDS-fragments.

4.1. Evaluation score for initial modification

The initial modification consisted of pitch-reduction of 23%. In addition to this prosodic modification, the keyword variants in the ASR-output were morphologically normalized as well. However, these modifications showed deterioration in all evaluation scores of the ASR-system (table 7). Compared to the baseline scores, precision remained 100%, recall decreased by 21.7% and F-score decreased by 20.8%.

Modification	Magnitude of modification	Precision	Recall	F-score
None (baseline)	----	*100.0%*	*55.4%*	*71.3%*
Pitch	-23%	100.0%	33.7%	50.7%

Table 7. *Initial evaluation scores for ASR-performance on modified IDS-fragments.*

4.2. Evaluation scores for further modifications

After the initial modification, the second set of modifications based on the proportions in table 5 were applied. These proportions were calculated at the hand of the average pitch and speaking rate of IDS and ADS measured in Mengru Han (2019), so that the IDS-fragments could be modified to resemble ADS' prosodic measures. The modifications also included the morphological normalization of keyword variants. Nevertheless, none of these modifications resulted in an improvement over the baseline results (table 8). The pitch-modification of -6% did not show such severe deterioration as the previous modification. Compared to the baseline scores, precision decreased by 0.6%, recall decreased by 4.3% and F-score decreased by 3.8%. The decrease in precision was caused by one false positive of the keyword "appel". Out of all modifications, the modification of +11% to the speaking rate showed the *least negative* results: precision remained 100%, recall decreased by 2.4% and the F-score by 2.0%. Lastly, a combination of these two modifications was applied. The precision remained 100%, while the recall decreased by 4.2% and the F-score by 3.6%.

Modification	Magnitude of modification	Precision	Recall	F-score
None (baseline)	----	*100.0%*	*55.4%*	*71.3%*
Pitch	-6%	99.4%	51.1%	67.5%
Speaking rate	+11%	100%	53.0%	69.3%
Pitch and speaking rate	-6% and +11%	100%	51.2%	67.7%

Table 8. *Evaluation scores for ASR-performance on modified IDS-fragments. Baseline scores are in cursive.*

4.3. Overview of all modifications and corresponding evaluation scores

To compare the evaluation scores resulting from all modifications, all F-scores are visually represented in figure 1. The y-axis represents the F-score. On the x-axis, the baseline and all four prosodic modifications are represented.

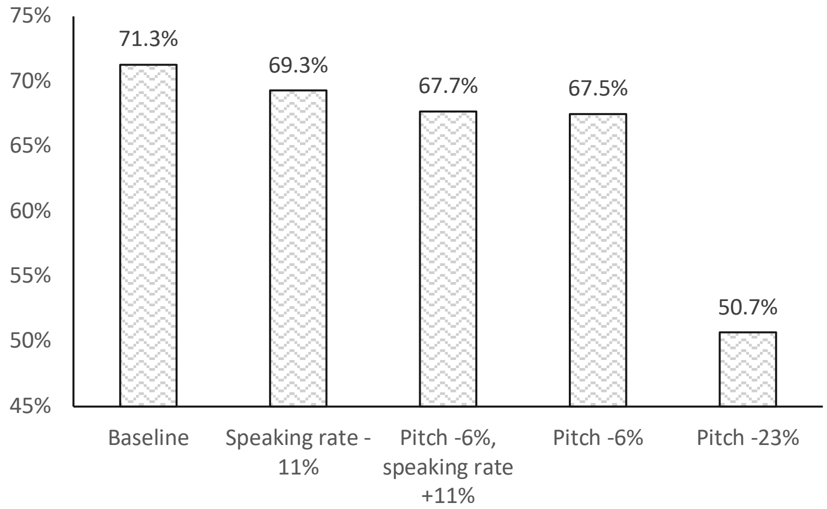

Figure 1. *Comparison of F-scores for ASR-performance (on the y-axis) for the baseline (Van der Klis et al., 2020) and all modifications (on the x-axis).*

5. Discussion

In this study, it was found that prosodic modifications of IDS did not have positive effects on the performance of ASR. In the baseline study of Van der Klis et al. (2020), ASR-performance was measured for IDS-fragments for 18-month-olds. The resulting F-score was 71.3%, with a precision of 100.0% and a recall of 55.4%. These same fragments were prosodically modified in the current study, which aimed to improve these scores. However, no improvements were found. The best evaluation scores that were acquired were an F-score of 69.3%, with a precision of 100% and a recall of 53.0%. The remaining modifications all led to worse results, as can be seen in figure 1. This study's research questions asked whether the automatic recognition of IDS could be improved by adjusting the prosodic properties of IDS before feeding it to the ASR-system. The hypothesis stated that such adjustments would improve the performance of ASR for IDS. Unfortunately, neither modifications applied to pitch, speaking rate, or pitch and speaking rate led to results that could support the hypothesis. This outcome is somewhat unexpected, since it could not replicate the positive results of previous studies concerning ASR for CS (Gustafson & Sjölander, 2002; Stemmer et al., 2003), even though the prosodic properties of this register strongly overlap with IDS.

5.1. Limitations of the current study and inspiration for future research

The current study did have its limitations. First of all, the high variability of the realisation of IDS between speakers was not taken into account. Only the entire sets of speakers' fragments were modified and evaluated. The results do therefore not reflect possible positive effects that the modifications might have had on the performance of ASR per speaker. Furthermore, because of the scope of this study, a limited number of modifications was applied to the speech-fragments. These limitations and some additional points of interest are discussed below.

First of all, as shown by previous studies' results, IDS shows a higher degree of variability than ADS (Fernald & Simon, 1984; Fernald et al. 1989; Cristia & Seidl, 2014;

Miyazawa et al., 2017). Specifically, Cristia and Seidl (2014) and Miyazawa et al. (2017) illustrate this variability in acoustic terms. In their studies, the vowel shows an increase in IDS compared to ADS, which generally indicates that the range of the first and second formant is extended beyond the range typical of ADS. In Fernald and Simon (1984) and Fernald et al. (1989), the same trend is illustrated in terms of prosody, where the average pitch-range of sentences uttered is greater for IDS compared to ADS. Since ASR relies on these properties of speech for its performance, a relatively high variability might make it harder for such a system to perform well. If this would be the case, then modifying IDS for its pitch and speaking rate would leave performance relatively unaffected, since the higher variability of IDS remains.

Secondly, modifications of IDS-fragments as a whole might cause a decrease in ASR-performance. That is because higher average variability of IDS does not necessarily mean that utterances show extreme prosodic outliers at every point in time. Some segments might even have ADS-like properties. Thus, for prosodic modifications of IDS to have a positive effect, it might be more useful to only apply these modifications on the segments of IDS that show actual prosodic outliers typical of IDS, instead of applying them on speech-fragments as a whole. This would leave the ADS-like segments unmodified, which might contribute to a better ASR-performance.

Finally, in the same vein as the previous point, looking at the realization of IDS per speaker, or per speech-fragment, could yield more detailed insights into how the prosodic properties actually diverge from ADS. The exact individual properties of spoken language differ between speakers and they all have their very own prosodic idiosyncrasies. By using a single prosodic average from different speakers to determine prosodic modifications, such personal properties of IDS might be levelled out. For future research, this would mean that determining and applying the necessary prosodic modifications per speaker could yield better performance for ASR.

6. Conclusion

IDS shows remarkable prosodic differences with ADS. These differences include a higher average pitch, a lower speaking rate and higher variability, and might cause problems for the automatic recognition of IDS. This thesis discussed the properties of IDS and posited the main research question: can ASR's performance on IDS be improved by adjustment of its prosodic properties? The hypothesis was that the modification of IDS would improve ASR's performance. However, the findings did not support this hypothesis. Although the results did not provide any direct means for improving ASR for IDS, this study does provide insights for future research on the matter and brings us one step closer to better automatic recognition for IDS.

Several points of interest for future research were brought forward in the discussion. First, this study only tested three types of modifications. This still leaves many more different modifications and their combinations to be tested in the future. Second, IDS' high variability could be an important factor in ASR's performance. Factoring in this variability in prosodic adjustments of speech might therefore lead to more positive effects. Third, adjusting speech on the word level instead of on a more general level might yield better improvements for automatic recognition. Finally, future research might benefit from a more focused approach, where modifications are applied based on individual speakers' speech instead of on the average of a set of speakers.

Bibliography

Audacity (3.0.2.). (2021). [Audio software]. Audacity Team. Retrieved from https://www.audacityteam.org

Benders, T. (2013). Mommy is only happy! Dutch mothers' realisation of speech sounds in infant-directed speech expresses emotion, not didactic intent. *Infant Behavior and Development*, *36*(4), 847–862. https://doi.org/10.1016/j.infbeh.2013.09.001

Cooper, R. P., & Aslin, R. N. (1990). Preference for Infant-Directed Speech in the First Month after Birth. *Child Development*, *61*(5), 1584–1595. https://doi.org/10.2307/1130766

Cristia, A., & Seidl, A. (2014). The hyperarticulation hypothesis of infant-directed speech. *Journal of Child Language*, *41*(4), 913–934. https://doi.org/10.1017/s0305000912000669

Fernald, A., & Simon, T. (1984). Expanded intonation contours in mothers' speech to newborns. *Developmental Psychology*, *20*(1), 104–113. https://doi.org/10.1037/0012-1649.20.1.104

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. https://doi.org/10.1017/s0305000900010679

Ghai, S., & Sinha, R. (2015). Pitch adaptive MFCC features for improving children's mismatched ASR. *International Journal of Speech Technology*, *18*(3), 489–503. https://doi.org/10.1007/s10772-015-9291-7

Gustafson, J., & Sjölander, K. (2002, September). *Voice Transformation for Children's Speech Recognition in a Publicly Available Dialogue System*. 7th International Conference on Spoken Language Processing, Denver, Colorado.

Han, M. (2019). *The Role of Prosodic Input in Word Learning.* [Dissertation, Utrecht University]. Published by LOT. ISBN: 978-94-6093-319-6

Hagen, A., Pellom, B., & Cole, R. (2007). Highly accurate children's speech recognition for interactive reading tutors using subword units. *Speech Communication*, *49*(12), 861–873. https://doi.org/10.1016/j.specom.2007.05.004

Jurafsky, D., & Martin, J. H. (n.d.). Evaluation: Precision, Recall, F-Measure. In *Speech and Language Processing* (3rd draft ed., pp. 11–13).

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. https://doi.org/10.1121/1.1869172

Kitamura, C., Thanavishuth, C., Burnham, D., & Luksaneeyanawin, S. (2002). Universality and specificity in infant-directed speech: Pitch modifications as a function of infant age and sex in a tonal and non-tonal language. *Infant Behavior and Development*, *24*(4), 372–392. https://doi.org/10.1016/s0163-6383(02)00086-3

Van der Klis, A., Adriaans, F., Han, M., & Kager, R. (2020). Automatic Recognition of Target Words in Infant-Directed Speech. *Companion Publication of the 2020 International Conference on Multimodal Interaction*, 522. https://doi.org/10.1145/3395035.3425184

Miyazawa, K., Shinya, T., Martin, A., Kikuchi, H., & Mazuka, R. (2017). Vowels in infant-directed speech: More breathy and more variable, but not clearer. *Cognition*, *166*, 84–93. https://doi.org/10.1016/j.cognition.2017.05.003

Nederlandse Taalunie. (1998–2004). *Corpus Gesproken Nederlands*. [Dataset].

Potamianos, A., Narayanan, S., & Lee, S. (1997, September). *Automatic Speech Recognition for Children*. 5th European Conference on Speech Communication and Technology, Rhodes, Greece.

Povey, D., Ghoshal, A., Boulliane, G., Burget, L., Glembek, O., Goel, N., Hanneman, M., Motlicek, P., Qian, Y., Schwarz, P., Silovsky, J., Stemmer, G., & Vesely, K. (2011). *The Kaldi Speech Recognition Toolkit* [Automatic speech recognition software]. IEEE Signal Processing Society. http://kaldi-asr.org

Soderstrom, M. (2007). Beyond babytalk: Re-evaluating the nature and content of speech input to preverbal infants. *Developmental Review*, *27*(4), 501–532. https://doi.org/10.1016/j.dr.2007.06.002

Stemmer, G., Hacker, C., Steidl, S., & Nöth, E. (2003, September). *Acoustic Normalization of Children's Speech*. EUROSPEECH 2003 - 8th European Conference on Speech Communication and Technology, Geneva, Switzerland.

Stichting Open Spraaktechnologie. (2019, May 10th). Stichting Open Spraaktechnologie. https://openspraaktechnologie.org