Multimodal Technologies and Interaction

MDPI

Article

Using Open-Source Automatic Speech Recognition Tools for the Annotation of Dutch Infant-Directed Speech

Anika van der Klis [1,*], Frans Adriaans [1,*], Mengru Han [2] and René Kager [1]

[1] Institute for Language Sciences, Utrecht University, 3512 JK Utrecht, The Netherlands; r.w.j.kager@uu.nl
[2] Department of Chinese Language and Literature, East China Normal University, Shanghai 200241, China; mrhan@zhwx.ecnu.edu.cn
* Correspondence: a.vanderklis@uu.nl (A.v.d.K.); f.w.adriaans@uu.nl (F.A.)

Abstract: There is a large interest in the annotation of speech addressed to infants. Infant-directed speech (IDS) has acoustic properties that might pose a challenge to automatic speech recognition (ASR) tools developed for adult-directed speech (ADS). While ASR tools could potentially speed up the annotation process, their effectiveness on this speech register is currently unknown. In this study, we assessed to what extent open-source ASR tools can successfully transcribe IDS. We used speech data from 21 Dutch mothers reading picture books containing target words to their 18- and 24-month-old children (IDS) and the experimenter (ADS). In Experiment 1, we examined how the ASR tool Kaldi-NL performs at annotating target words in IDS vs. ADS. We found that Kaldi-NL only found 55.8% of target words in IDS, while it annotated 66.8% correctly in ADS. In Experiment 2, we aimed to assess the difficulties in annotating IDS more broadly by transcribing all IDS utterances manually and comparing the word error rates (WERs) of two different ASR systems: Kaldi-NL and WhisperX. We found that WhisperX performs significantly better than Kaldi-NL. While there is much room for improvement, the results show that automatic transcriptions provide a promising starting point for researchers who have to transcribe a large amount of speech directed at infants.

Keywords: infant-directed speech; automatic speech recognition; research tools; speech registers; transcriptions

check for updates

Citation: van der Klis, A.; Adriaans, F.; Han, M.; Kager, R. Using Open-Source Automatic Speech Recognition Tools for the Annotation of Dutch Infant-Directed Speech. *Multimodal Technol. Interact.* **2023**, *7*, 68. https://doi.org/10.3390/mti7070068

Academic Editor: Alexey Karpov

Received: 24 May 2023
Revised: 23 June 2023
Accepted: 26 June 2023
Published: 3 July 2023

Copyright: © 2023 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (https://creativecommons.org/licenses/by/4.0/).

1. Introduction

When addressing infants, adults spontaneously adopt a different speech register referred to as infant-directed speech (IDS) or baby talk [1–3]. This speech register is characterised by a variety of intonational and prosodic characteristics, including a higher mean pitch, a larger pitch range, and greater pitch variability compared to adult-directed speech (ADS) (for a review, see [4]). IDS has also been found to have a slower speaking rate than ADS in many languages, including Dutch [5,6]. Many studies have reported positive links between the acoustic properties of IDS and children's linguistic outcomes (for a meta-analysis, see [7]). The mechanisms driving this relationship are still widely debated.

Previous studies have shown that slow speech improves children's word recognition performance [8,9]. Han et al. showed that Dutch mothers slowed down speech when introducing unfamiliar words compared to familiar words [6]. The results are less conclusive for pitch. In Singh et al., 7- and 8-month-old infants were able to recognise words that were previously presented in IDS but not when they were presented in ADS [10]. Similarly, Estes and Hurley showed that 17.5-month-old children only learned novel words in IDS but not in ADS [11]. The effects of pitch have not been studied in isolation; thus, it remains unclear whether the facilitative effects of pitch on word recognition can be attributed to pitch alone. In addition, Han et al. found that Dutch mothers increase pitch for familiar words, while Chinese mothers increase pitch for unfamiliar words [12]. Pitch may function differently in these languages, and it remains unclear how pitch facilitates learning. It has also been suggested that infants prefer listening to IDS over ADS [13–15], indirectly facilitating the learning process.

The exaggerated prosody of IDS may also facilitate the learning of vowel categories. Kuhl et al. suggested that vowels in IDS are acoustically more extreme, containing larger vowel spaces than vowels produced in ADS, leading to this hypothesis [3]. Recent studies have reported that IDS vowels are produced with higher variability compared to vowels in ADS [16,17], resulting in more overlap between vowel categories in IDS. Adriaans and Swingley proposed that among this high variability, however, mothers produce exaggerated high-quality instances of vowels that can facilitate vowel categorisation [18]. The trade-off between larger vowel spaces, higher variability, smaller contrasts, and the presence of high-quality tokens in the input remains to be seen.

There seems to be a general trend of IDS becoming prosodically more like ADS as children grow older [12,19,20]. Specifically, Han et al. found that utterance mean pitch was significantly lower when Dutch mothers addressed their 24-month-old infants compared to addressing their 18-month-old infants, even though utterance mean pitch was still higher in IDS than ADS at both ages [12]. Sjons et al. found an increase in articulation rate of Swedish IDS from 7 to 33 months, suggesting that the articulation rate of IDS becomes more similar to the articulation rate of ADS as children grow older. Nevertheless, IDS was still slower than ADS [20]. Han et al., on the other hand, did not find any evidence of age-related changes in articulation rate in Dutch IDS from 18 to 24 months, and speaking rate remained slower compared to ADS [6]. Age-related effects vary cross-linguistically, but there is a trend of IDS becoming acoustically more like ADS over time.

To identify and analyse the distinctive properties of IDS and subsequently advance our understanding of the role of IDS in language development, it is essential to collect and transcribe IDS in many languages and across many speakers for infants at different ages. Preparing speech data for analysis takes a notoriously long time. Segmenting, annotating, and transcribing an hour of speech, including verifying the quality of the transcription, can take up to fifty hours in total depending on the contents [21]. Depending on the research aims and the accuracy needed to accomplish these, automatic transcriptions may be used as a starting point and then manually corrected by a human annotator to save time [22]. Tools are being developed to generate automatic annotations that would benefit research on IDS by speeding up the annotation process [23]. Currently, it is still common practice in the field to transcribe speech manually. To date, studies have not yet addressed to what extent we can use off-the-shelf ASR tools to facilitate the annotation process of IDS. The current study assessed the performance of ASR tools in the annotation of Dutch speech directed at 18-month-old and 24-month-old infants. In Experiment 1, we examined how the ASR tool Kaldi-NL performs at annotating target words in IDS vs. ADS. In Experiment 2, we examined the performance more broadly by testing two different ASR systems (Kaldi-NL and WhisperX) on the complete set of IDS utterances. We compared their performance in terms of word error rates (WERs). The experiments inform us to what extent off-the-shelf ASR tools trained on ADS are successful at annotating Dutch IDS.

2. Previous Work

Automatic speech recognition (ASR) is the process of generating text representations for acoustic speech input. ASR systems have components that require extensive training, such as an acoustic model and a language model. The acoustic model learns from audio recordings combined with phonetic transcriptions, creating statistical representations of speech sounds. Many ASR systems use deep neural networks to create these representations, drastically improving their automatic transcription performance [24]. The acoustic model translates the audio signal into a sequence of the most probable phonemes. The language model learns from a large corpus of transcribed speech, creating statistical probabilities of word sequences in the language. The ASR system combines the two models to produce the most likely written transcription of the signal as output.

Previous studies have examined whether certain acoustic features are more likely to result in ASR errors. Fast speech and extremely long word durations are both related to higher error rates [25–27]. Goldwater et al. found that extreme values of pitch and intensity also increase error rates [25]. In addition, an analysis of two human–computer dialogue systems shows that misrecognised utterances are associated with pitch excursions, loudness, and longer duration [28]. The authors marked these as instances of hyperarticulated speech. Importantly, some of these features (i.e., above-average mean pitch and pitch range and below-average speaking rate) are similar to the typical features of IDS.

Precisely because IDS is a highly variable and exaggerated speech register, it has been hypothesised that it may also serve as particularly good training data, resulting in more robust models when the goal is to transcribe a less variable speech register such as ADS [29,30]. The acoustic characteristics of IDS—potentially resulting in phonetic categories that are well separated in the input space—could aid phonetic classification using Gaussian mixture models. Kirchhoff and Schimmel trained an ASR system on IDS using recordings of 22 American English mothers addressing their 2- to 5- month-old infants, and they trained another system on ADS using the same mothers addressing the adult experimenter. The system trained on ADS was highly accurate at recognising target words in ADS (95.5%) but less in IDS (81.6%). The system trained on IDS was notably better at recognising target words in IDS (93.5%), but it did not perform as well on recognising the same target words in ADS (90.2%) [29]. These results indicate that a matched system (i.e., trained and tested on the same speech register) produces the best recognition results. The largest degradation in performance is found when an ASR system trained on ADS is used for the recognition of IDS which is the more variable speech register. Nevertheless, the authors used a relatively small set of training data (utterances by 22 speakers). Currently, we do not know whether an ASR system trained on a much larger ADS data set contains more robust models that are more suitable for the recognition of IDS.

3. Experiment 1

In the first experiment, we aimed to assess to what extent an open-source ASR tool, Kaldi-NL, was successful at annotating target words in continuous, semi-naturalistic IDS. This is the first study to (1) address this question for Dutch, (2) use a readily available open-source ASR tool, (3) compare the recognition performance of IDS addressed to different age groups, and (4) examine the effects of different acoustic features (i.e., mean pitch, pitch range, and articulation rate) on recognition accuracy. While acoustic deviations in pitch and speaking rate support children's word recognition abilities [8–11], previous studies have shown that these may have negative effects on ASR performance [25–27]. Given that ASR performance is negatively affected by acoustic deviations, and that Dutch IDS is marked by a higher mean pitch, a larger pitch range, and a slower articulation rate compared to Dutch ADS, we would expect that an ASR system trained on Dutch ADS exhibits lower performance when transcribing Dutch IDS. Very few studies have assessed the accuracy of ASR systems at transcribing IDS, and there are none so far for Dutch. It is important to verify whether research findings generalise to other languages. First, we compared the accuracy of Kaldi-NL at transcribing target words produced by Dutch mothers embedded in continuous IDS directed at 18-month-old children and 24-month-old children and the same target words embedded in continuous ADS directed at the experimenter. Then, we examined which acoustic features affected speech recognition accuracy using a logistic mixed-effects model. The results informed us to what extent an off-the-shelf ASR tool can successfully transcribe IDS and whether the transcription accuracy is negatively affected by IDS.

3.1. Materials and Methods

3.1.1. Participants

This study is part of a larger cross-linguistic corpus of Dutch and Mandarin Chinese infant-directed speech [31]. The speech data collection methods are identical to those reported in [6,12]. From this corpus, we included 21 Dutch-speaking mother–child dyads who were recruited from the Utrecht Baby Lab database and were all Dutch native speakers living in the Utrecht area in the Netherlands. We used a longitudinal design and collected mothers' ADS and IDS speech data when their children were 18 months old (*M* = 18.4, range = 18.0–18.9) and 24 months old (*M* = 24.7, range = 24.0–27.0). All mothers were native speakers of Dutch with higher education (undergraduate degree and above). All children were typically developing with no reports of language or hearing problems. All participating mothers signed informed consent forms.

3.1.2. Materials and Procedure

Mothers read the same picture book to their infant to elicit IDS and to the female adult experimenter to elicit ADS during the recording sessions. Different picture books for each time point (children's ages of 18 months and 24 months) were designed to elicit two different sets of seven disyllabic target words. On each page, the target word was on the left side and an illustration including a depiction of the word was on the right side (for the picture book, see [31], p. 187). The mothers were instructed to tell the story including the target words, eliciting semi-naturalistic speech. The target words at both time points can be found in Table 1. These target words were selected because they were likely unfamiliar to the child (apart from "apple" and "grandpa", which were used for comparison), which was relevant to the previous study. The unfamiliar target words at 24 months are of much lower frequency than the unfamiliar target words at 18 months.

In total, the participants produced 1051 target words embedded in semi-naturalistic speech across both speech registers and time points. All mothers produced each target word embedded in utterances at least once in each condition at each age. The productions are equally distributed: 563 target word productions when the infants were 18 months old (243 in ADS; 320 in IDS) and 488 target word productions when the infants were 24 months old (215 in ADS; 273 in IDS). The total duration of the speech sample was 97.95 min (ADS: 36.48 min; IDS: 61.47 min) at 18 months and 102.35 minutes (ADS: 35.65 min; IDS: 66.70 min) at 24 months. All participants were tested in a quiet room in the Utrecht Baby Lab. The audio recordings were made using a ZOOM H1 recorder with 16-bit resolution and a sampling rate of 44.1 kHz.

Table 1. Target words and their word frequencies according to the SUBTLEX corpus of Dutch [32].

18 Months			**24 Months**		
Dutch	**Translation**	**Frequency**	**Dutch**	**Translation**	**Frequency**
opa	"grandpa"	2507	opa	"grandpa"	2507
appel	"apple"	446	appel	"apple"	446
eland	"moose"	115	emoe	"emu"	6
bever	"beaver"	128	wezel	"weasel"	90
walnoot	"walnut"	31	bamboe	"bamboo"	30
kasteel	"castle"	1207	kapel	"chapel"	194
pompoen	"pumpkin"	109	jasmijn	"jasmine"	37

3.1.3. Transcriptions

We compared the automatic annotations to the manual annotations of target words to assess the accuracy of the ASR system at annotating target words in IDS. All target words were manually annotated in previous work (for details, see [31]). A trained Dutch native speaker extracted the target words from the audio recordings using Praat [33]. For the current study, the full recordings were automatically transcribed using the online Kaldi-NL ASR tool developed by the Dutch Foundation of Open Speech Technology and hosted by the Radboud University (Version 0.5.0; [34]). The Dutch models were developed at the University of Twente using the Spoken Dutch Corpus ("Corpus Gesproken Nederlands") containing about 900 h of Dutch speech recordings from, for example, conversations and television shows [35]. Kaldi-NL has a lexicon of ca. 250 thousand words and employs time-delay neural network layers, which have been shown to outperform low-frame-rate bidirectional long short-term memory acoustic models [36]. A recent study used Kaldi-NL to transcribe Dutch doctor–patient consultation recordings and initially found a WER of 25.8% without fine-tuning the language model or lexicon to include domain-specific healthcare words [37]. To generate the automatic transcriptions, the online system takes audio files (e.g., WAV) as input. After a short period of processing, the output of the ASR system consists of a plain text file containing a written transcription and a CTM file containing all transcribed words and their corresponding timestamps (i.e., indicating when they occurred in the audio file). Using this output, we lastly examined the accuracy of the automatic annotations of target words using the evaluation procedure described below.

3.1.4. Evaluation Procedure

We compared the automatic annotations of the target words from the time-stamped CTM files to the manual annotations (i.e., the ground truth) using an interactive Python script. For each target word in the manual annotations, the script shows the timestamp of the word from the Praat TextGrid and the timestamp of the target word from the automatic transcription in addition to playing both extractions from the audio file. The evaluator checks whether the manual and automatic words match. If yes, this counts as one "hit" (correctly identified target word). If not, then the word is a "miss" (not identified target word). Lastly, the script collects all target words that were automatically transcribed but not matched to a manual annotation and marks these as "false positives" (words incorrectly identified as target words). These cases were double-checked since the target words could potentially have been overlooked during the manual annotation process. The output of the script is a data file containing all assessed target words, the speech register (IDS or ADS), the time point (18 m or 24 m), the assessment (hit, miss, or false positive), and the timestamps from the TextGrid and from Kaldi-NL. All morphological varieties of the target words, such as diminutives (e.g., *appeltje* or *walnootje*) were also included in the data. We only analysed target words and not full sentences because it is easier to compare the data across speech registers and to eliminate the chance that any observed differences between IDS and ADS can be attributed to the language model (e.g., IDS tends to have shorter sentences and more repetitions) or the vocabulary size (e.g., IDS tends to have shorter, simplified words).

The frequencies of hits, misses, and false positives allow us to calculate three common accuracy scores: recall, precision, and *F*-scores. Recall informs us how many of the total target words annotated manually were also found by the ASR system. Precision informs us how many of the recalls were target words and not false positives. *F*-score is the harmonic mean between recall and precision [38]. This is an important additional measure because high recall does not equal high accuracy when precision is low, and vice versa. The measures are calculated as follows:

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

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

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

3.1.5. Acoustic Features of Target Words

We examined each target word's mean pitch, pitch range, and articulation rate. First, we automatically extracted the minimum pitch, maximum pitch, and mean pitch from each target word in IDS and ADS using a pitch range of 100–600 Hz in Praat version 6.1.09 [33]. The top and bottom 5% of pitch measurements were all manually checked for pitch jumps (i.e., halving or doubling). In the case of a pitch jump, the pitch range was slightly adjusted to better fit the data. The pitch range was calculated by subtracting the minimum pitch from the maximum pitch of a target word. The articulation rate was calculated by dividing the number of syllables by the total duration of the target word in seconds (i.e., number of syllables per second). Target words were excluded when pitch could not be measured due to interference of the child's voice (n = 17) or due to whispering (n = 2), resulting in a final set of 1032 target words for the acoustic analysis.

3.1.6. Statistical Analysis

To assess what affects ASR performance, we examined the effects of speech register, infant age, and the different acoustic properties of IDS—mean pitch, pitch range, and articulation rate—on recognition accuracy. The results of 1032 target words were analysed by fitting logistic mixed-effects models using the *lme4* package version 1.1-30 [39] in *R* version 4.2.0 [40] to predict recognition accuracy for each target word (hit or miss). The continuous variables F_0 mean, F_0 range, and articulation rate were centred and scaled. We used dummy coding for the dichotomous variables speech register (IDS or ADS) and age (18 m or 24 m) with ADS and 18 m as reference levels. We added random intercepts for participants to account for potential individual variations in speech perceptibility and items because the target words were not the same across both time points and differed in word frequency. This can negatively impact recognition performance in a way that is not related to the measures that are of interest in the present study. Lastly, we calculated odds ratios from the regression coefficients to examine the impact of the predictors.

3.2. Results

We first calculated recall, precision, and F-scores to assess the accuracy of the ASR system for each speech register at each time point. This allowed us to compare the recognition accuracy for IDS to ADS. Then, we examined the distributions of the various acoustic measures across all conditions to examine whether the acoustic features that are typical of IDS—mean pitch, pitch range, and articulation rate—affected recognition accuracy. Lastly, we fitted a logistic mixed-effects model to examine which of the predictors has a significant effect on recognition accuracy.

3.2.1. Accuracy Scores

For speech addressed to 18-month-old infants, the ASR system correctly annotated 180 of 320 (56.3%) target words. For ADS, the ASR system found 174 of 243 (71.6%) target words. For the 24-month-old infants, the system correctly annotated 151 of 273 (55.3%) target words in IDS and 132 of 215 (61.4%) target words in ADS. The difference in recognition accuracy between ADS and IDS diminished between the two time points. The recall scores are visualised in Figure 1. All target words were correctly annotated at least once, indicating that none of the target words were out-of-vocabulary words (i.e., all target words are present in Kaldi-NL's vocabulary).

In both registers, precision is 100%. Precision is calculated using false positives, and there were none in the data. For false positives to occur, other produced words must be phonologically similar to target words, which is unlikely given the limited contents of the picture books used in the present study. Table 2 contains the results of the evaluation procedure.

Table 2. Results of the evaluation procedure in proportions.

	18 Months		**24 Months**	
Register	**ADS**	**IDS**	**ADS**	**IDS**
Recall	0.72	0.56	0.61	0.55
Precision	1.00	1.00	1.00	1.00
F-score	0.84	0.72	0.76	0.71

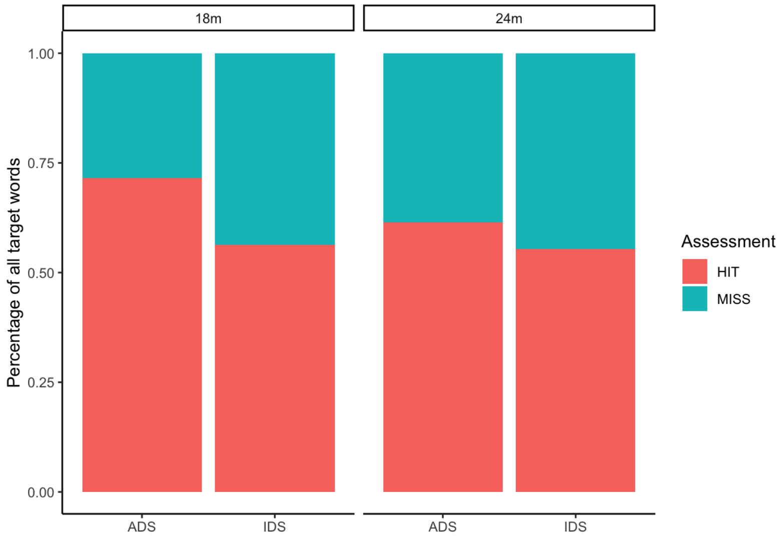

Figure 1. The proportions of hits and misses for each speech register within each age group.

Recall scores are generally lower for target words at 24 months and also for ADS. This is likely caused by the lower word frequencies of the target words that were produced at this age, as shown in Table 1. Low-frequency words have low probabilities in the language model of the ASR tool, making them less likely candidates to be selected. Therefore, the general word frequencies of the target words affect recognition accuracy. The important finding is that the decrease in recognition accuracy found for IDS compared to ADS became much smaller.

3.2.2. Acoustic Measures

Figure 2 shows boxplots of the mean pitch of hits and misses in both speech registers. First, the boxplots show that on average, target words in IDS have a higher mean pitch than target words in ADS at both time points. Target words have the highest mean pitch in IDS at 18 months. Secondly, missed target words have on average a higher mean pitch than hits at 18 months. This difference seems to have disappeared at 24 months, although missed target words seem to have more extreme mean pitch values in both directions.

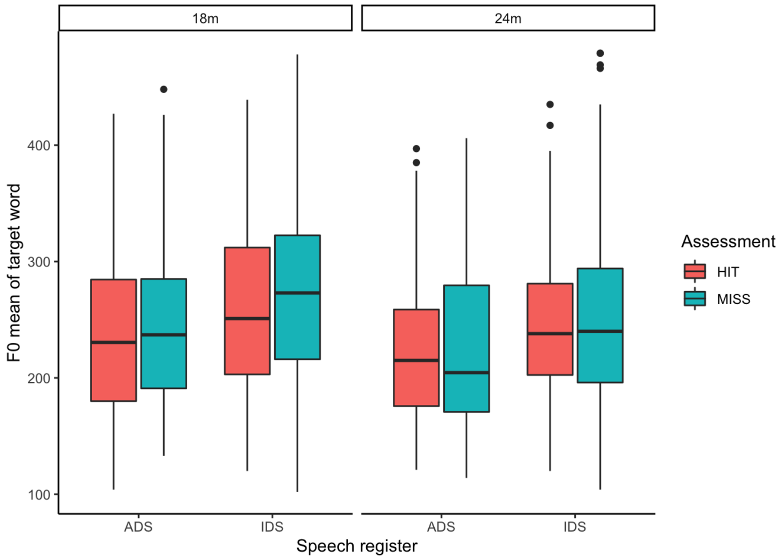

Figure 2. Boxplots of the mean pitch of target words.

Figure 3 shows boxplots depicting the average pitch ranges of target words. First, the boxplots show that target words in IDS have a larger pitch range on average compared to target words in ADS. The figure does not provide evidence that missed target words have larger pitch ranges than hits on average.

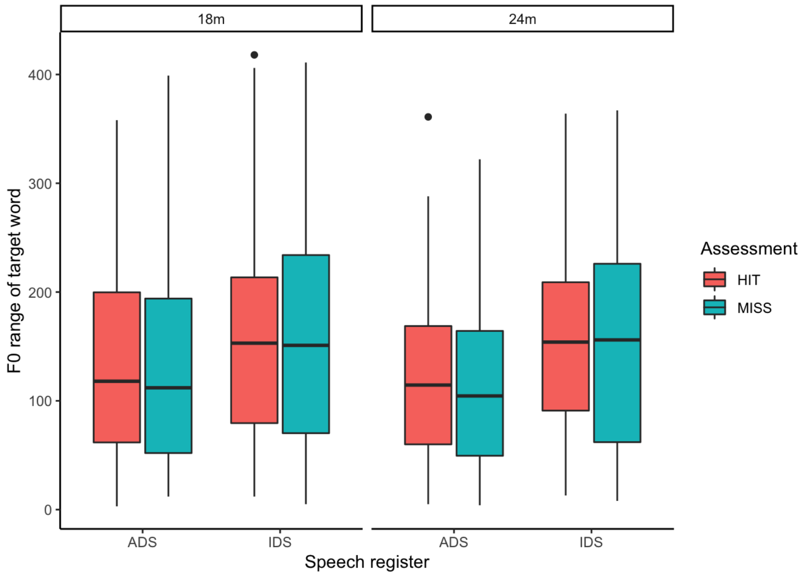

Figure 3. Boxplots of the pitch range of target words.

Figure 4 shows boxplots depicting articulation rates. First, target words at 24 months are on average produced faster than target words at 18 months. At 24 months, target words in IDS were produced slower than target words in ADS. The difference between IDS and ADS is surprisingly smaller at 18 months, whereas we would expect IDS to become more similar to ADS over time. One explanation could be that the target words are of much lower frequency at 24 months, and mothers may lower their articulation rates more for didactic purposes when presenting unfamiliar words to their children [6]. Articulation rate does not seem to have a large effect on recognition accuracy, although we find more extreme values of low articulation rates across missed target words.

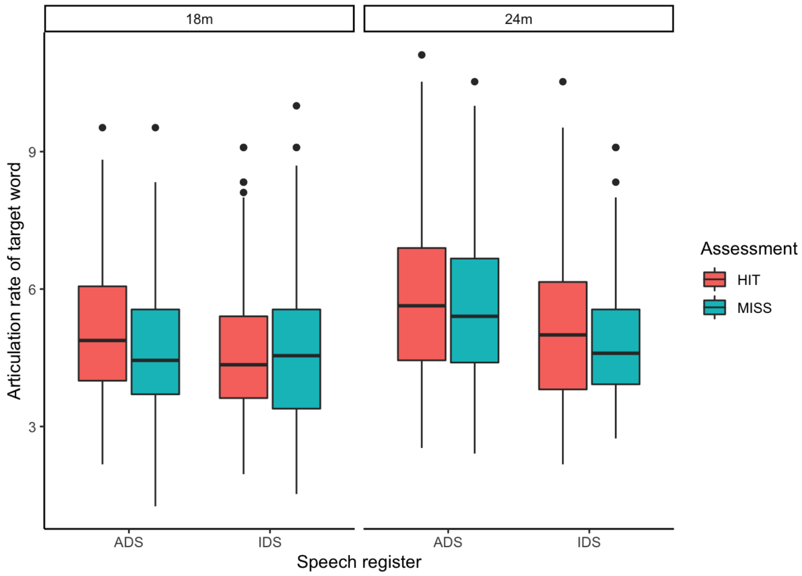

Figure 4. Boxplots of the articulation rate (syllables/s) of target words.

The results of 1032 target words were analysed by fitting logistic mixed-effects models using a bottom-up approach. First, we found that adding random intercepts for participants and items significantly improved model fit. This indicates there is random variability across participants and items that affects recognition accuracy. Then, we examined which of the fixed effects (speech register, age, mean pitch, pitch range, and articulation rate) significantly improved model fit. We found that speech register ($p < 0.001$) and mean pitch ($p = 0.02$) significantly improved the fit of the model. There is a significant effect of age if we do not add a random intercept for items to the model. The effect of age likely disappears when adding a random intercept for items since the items differed across the two measurement waves, partially accounting for this effect. The model that includes a random intercept for items better fits the data. There was no improvement to model fit when adding an interaction between the fixed effects. The final model, including fixed effects for speech register and mean pitch and random intercepts for participants and items, is presented in Table 3.

Table 3. Results of the logistic mixed-effects model transformed to exponentiated coefficients (accuracy $\sim$ speech register + F_0 mean + (1 | subject) + (1 | item)).

Predictor	Exp. Coefficient	*SE*	*Z*-Value	*p*-Value
(Intercept)	0.40	0.37	−2.51	0.01
Register (IDS)	1.86	0.15	4.08	<0.001
F_0 mean	1.20	0.08	2.24	0.02

The speech register IDS is a strong predictor of a recognition error (i.e., a missed target word). When the target word is produced in IDS, there is an increase of 1.86 (95% CI [1.38, 2.51]) in the odds of the ASR system missing the target word compared to a target word produced in ADS. On top of this, there is a significant negative effect of mean pitch on recognition accuracy. A one-unit increase in mean pitch results in an increase of 1.20 (95% CI [1.02, 1.40]) in the odds of the ASR system missing a target word. Words produced with a higher mean pitch are problematic for the recognition of target words in continuous speech.

3.3. Discussion

The results show that there is a large gap between the recognition accuracy of ADS and IDS, especially for speech directed at younger infants. Previous studies on IDS have shown that the acoustic features of IDS become less salient over time [12,19]. This could explain why the difference between IDS and ADS automatic recognition accuracy became smaller for speech addressed to 24-month-olds compared to 18-month-olds. As we expected based on a previous study on American English, the ASR tool trained on ADS is less successful at transcribing IDS than ADS [29]. The difficulties with transcribing IDS generalise to Dutch. While Kaldi-NL is trained on a significantly larger data set compared to the ASR system used in the previous study on American English IDS (i.e., 900 h of speech compared to a set of utterances by 22 speakers), this did not help much to overcome the difficulties with automatically recognising this speech register.

We also examined which factors are predictors of a recognition error made by Kaldi-NL. The results show that IDS as a speech register is an important predictor of a missed target word. We also found a significant negative effect of mean pitch on recognition accuracy. Previous studies found that slow speech facilitates word recognition in children [8,9], although it could hinder ASR performance [25–27]. We did not find a significant effect of articulation rate on ASR accuracy. It could be possible that the ASR system trained on ADS does not have difficulties with the larger pitch range or slower articulation rate of IDS, but the ASR system does show a decrease in accuracy when transcribing target words with a higher mean pitch. The results of the mixed-effects model suggest that IDS is also likely to be more difficult to be recognised by the ASR system for reasons beyond the examined acoustic measures, for example, due to the high amount of acoustic variability or any syntactic differences.

3.4. Follow-Up Experiment

We found that Kaldi-NL transcribed approximately half of the target words correctly in IDS and approximately two-thirds in ADS. Based on this performance, we asked two follow-up questions. First, the target words, while important to the previous experiment, constitute only a relatively small portion of the total set of words in the data. Many studies investigating the acoustics of speech measure prosody at the utterance level. As such, a tool that recognises utterances can be very useful for research. One important question is, thus: To what extent do the results of target words generalise to the automatic transcription of utterances? Second, since we tested only one particular system in Experiment 1, the question is: To what extent is the performance reflective of ASR systems in general. That is, to what extent are the recognition results similar across different ASR systems? In Experiment 2, we tackle these two questions in parallel by manually transcribing all IDS utterances in the data set and comparing the two different systems (Kaldi-NL and the newly available open-source WhisperX) on their ability to transcribe these utterances as measured by WERs. By analysing full sentences instead of target words, we have over twenty times more IDS data, while the data are less affected by the large differences in target word frequencies across the two time points.

4. Experiment 2

In this second experiment, we first aimed to evaluate how open-source ASR systems perform at transcribing utterances in Dutch IDS. We compared WERs of utterances in IDS directed at 18-month-old and 24-month-old infants. We expected that WERs are lower in speech directed at older infants since IDS becomes prosodically more similar to ADS as children grow older [12,19]. The results for speech addressed to 18-month-old children and 24-month-old children in the previous experiment were difficult to compare because the target words examined at the two time points were of vastly different word frequencies—negatively influencing overall ASR performance. By calculating WERs of full utterances, we reduce the influence of target word frequencies on the results. In addition, previous studies examining correlations between the acoustics of IDS and children's language outcomes usually measure prosody at the utterance level instead of—or in addition to—the word level (e.g., [8,9,41]). As such, a tool that recognises utterances correctly can be very useful for research.

Based on the performance of Kaldi-NL in the previous experiment, the second aim was to assess whether an ASR system trained on a much larger, semi-supervised data set performs better at transcribing Dutch IDS. It might be possible that a larger training set results in more robust models that are more successful at transcribing a more variable speech register such as IDS. We compared the WERs of two different open-source ASR systems (Kaldi-NL and WhisperX) for the transcription of Dutch IDS. The second experiment informed us whether the results of target words in the first experiment generalise to full utterances and across different ASR systems.

4.1. Materials and Methods

4.1.1. Participants

In Experiment 2, we included the same 21 Dutch-speaking mother–infant dyads from the larger cross-linguistic corpus of Dutch and Mandarin Chinese infant-directed speech [31] that were used in the previous experiment.

4.1.2. Transcriptions

We used the same automatic transcriptions of the picture-book reading recordings that were described in the previous experiment generated by the open-source tool Kaldi-NL. Instead of only examining target words, however, we used the automatic transcriptions of all utterances in the recordings of the picture-book reading sessions (*M* = 348 words per recording). To calculate WERs, we manually annotated all words in the IDS recordings. A research assistant was trained to manually correct and supplement the Kaldi-NL transcriptions. All words except for the occasional mentions of children's names were included in the annotation process. Children's names were also removed from the automatic transcriptions. The manual annotation procedure resulted in a gold standard IDS data set containing a total of 15,309 words. All unique words and their total frequencies in the IDS corpus can be found on OSF (see the Data Availability Statement). Out of this total data set, only 4.4% of the words were target words. The influence of target words in this experiment is thus minimal, as are their potential frequency effects on the outcomes.

For the comparison between two open-source ASR systems, we also automatically transcribed the same IDS recordings using WhisperX [42], which provides improved accuracy and word-level timestamps using voice activity detection and forced phoneme alignment while using OpenAI's Whisper models [43]. Whisper contains weakly supervised (or semi-supervised) cross-linguistic training models (i.e., audio paired with unvalidated transcripts from the Internet), which allows for a larger quantity of training data compared to supervised models. A larger quantity of training data can result in more-robust models. The full data set comprises over 680,000 h of training data, of which 117,000 h cover 96 other languages. When testing the largest Whisper model on the Fleurs data set, a mean WER of 4.4% was found for English and a mean WER of 6.7% was found for Dutch [43]. WhisperX also takes audio files as input (e.g., WAV) and generates text files containing all transcribed words and their corresponding timestamps as output, which can be used in the evaluation procedure.

4.1.3. Evaluation Procedure

To calculate WERs, we used the toolkit *sclite* version 2.10 from SCTK version 2.4.12 [44], which is an open-source tool for scoring and evaluating the output of ASR systems. All reference and hypothesis transcription files were transformed to CTM format before being submitted to *sclite*. The tool calculates the WER in percentages for individual speakers by dividing the sum of word deletions, insertions, and substitutions by the total number of words in the human-labelled transcription. The higher the WERs, the lower the accuracy of the transcriptions. We standardised the texts by making all words lowercase and removing all punctuation in the ASR output and the reference transcriptions. In addition, common abbreviations were spelled out in full (e.g., *'m* => *hem*, *z'n* => *zijn*).

4.1.4. Statistical Analysis

In addition to reporting the overall WERs, a statistical analysis was carried out in *R* version 4.2.0 [40]. We fitted a linear mixed-effects model using the *lme*4 package version 1.1-30 [39] with WERs for each speaker as continuous outcome variables. Each speaker has four WER scores: two generated by Kaldi-NL and two generated by WhisperX, one for each measurement point. We included the ASR system (Kaldi-NL or WhisperX) and age (18 months and 24 months) as two dichotomous predictors to the model. We used dummy coding where Kaldi-NL and 18m were used as reference levels. We also added random intercepts for participants to the model. This allowed us to examine (1) whether WERs are affected by children's ages and (2) whether WERs are affected by the open-source ASR tool used to generate the transcriptions.

4.2. Results

Across both time points, Kaldi-NL had a mean WER of 40.12% (*SD* = 10.39). WhisperX had a mean WER of 22.49% (*SD* = 10.28). Table 4 presents descriptive results of WERs of Kaldi-NL and WhisperX for speech directed at 18-month-old and 24-month-old infants. There is a large difference in performance between Kaldi-NL and WhisperX but only small differences in performance between the two measurement waves.

Table 4. Descriptive statistics of WERs of transcriptions by Kaldi-NL and WhisperX for speech directed at 18-month-olds and 24-month-olds.

	18 Months		**24 Months**	
ASR System	*M*	*SD*	*M*	*SD*
Kaldi-NL	41.97	11.64	38.27	8.86
WhisperX	21.84	11.52	23.14	9.11

First, we compared the model including the predictor ASR system to a null model without any predictors. The model including the predictor ASR system provided a significantly better fit to the data compared to the null model ($p < 0.001$). Then, we added age to this model. The model including age did not provide a significantly better fit to the data ($p = 0.627$). The ASR systems did not perform differently on IDS directed at 18-month-olds or 24-month-olds. The results of the final model are shown in Table 5. The results show that the ASR system WhisperX significantly reduced WERs by 17.63% (95% CI [−21.52, −13.74]) on average compared to Kaldi-NL. The residuals in the model were normally distributed.

Table 5. Results of the linear mixed-effects model (WER∼ASR system + (1 | subject)).

Predictor	**Estimate**	***SE***	***t*-Value**
(Intercept)	40.12	1.78	22.52
System (WhisperX)	−17.63	1.97	−8.95

4.3. Discussion

Previous studies examining relations between the prosody of maternal speech and children's linguistic outcomes typically measured the acoustics at the utterance level [8,9,41]. As such, it is important that an ASR tool recognises utterances correctly. The results show that Kaldi-NL is not very successful at transcribing utterances in IDS. The mean WER of 40.12% is much higher than previously reported by Tejedor-García et al. for the healthcare domain [37]. They used Kaldi-NL to transcribe Dutch doctor–patient consultations and found a WER of 25.8%. We found that WhisperX performs significantly better. Although the mean WER of 22.49% is higher than the WER of 6.7% reported for the Fleurs data set using Whisper [43], WhisperX could be used as a starting point to facilitate the annotation process of IDS. The decrease in WER for both ASR systems compared to other Dutch ADS data sets corroborates the finding of Experiment 1 that IDS is more difficult to automatically transcribe compared to ADS, at least for systems trained on ADS.

The two ASR systems differ vastly in the amount of training data. Where the models of Kaldi-NL are trained on approximately 900 h of Dutch speech from television shows and lectures, Whisper is trained on 680,000 h of cross-linguistic training data (of which 117,000 h covers 96 other languages). We found that WhisperX performed significantly better than Kaldi-NL at transcribing full utterances in Dutch IDS. The large number of open-source ASR systems available can make it difficult for researchers to know which system best suits their needs. If multiple open-source ASR tools are available in a language, we would advise researchers to assess which system performs better on their specific data set. The results of this experiment show that the type of ASR system can have a large influence on the accuracy of the transcriptions.

5. General Discussion

In the current study, we aimed to assess to what extent open-source ASR tools can be used for the transcription of maternal speech directed at 18-month-old and 24-month-old infants. This is the first study to examine the transcription accuracy of IDS using off-the-shelf ASR tools trained on large, (semi-)supervised ADS data sets. Currently, most researchers of IDS transcribe audio recordings manually from scratch, while a growing number of open-source ASR tools trained on large data sets are available cross-linguistically.

Although the manual procedure results in highly accurate transcriptions, it is labour-intensive, which makes the data annotation process time-consuming and expensive. To date, no studies have examined whether researchers can successfully use off-the-shelf ASR tools trained on ADS for the annotation of IDS. Using automated tools can drastically decrease the time that is currently needed for manual transcriptions.

The results show that the open-source ASR system Kaldi-NL is less accurate when transcribing IDS compared to ADS. We found that the recognition accuracy of target words is decreased when they are produced in IDS compared to ADS, and we also found a negative effect of mean pitch. The difference in accuracy between the two speech registers was largest for speech directed at younger children. These results suggest that we first have to identify whether ASR tools can provide benefits before we start implementing them in the annotation process. A previous study found that WERs should be below 30% for automatic transcriptions to be beneficial to the annotation process [22]. Otherwise, it would be faster to annotate manually from scratch. Although we believe this limit can be different depending on the types of recognition errors or the specific research goals, the results of the current study constitute evidence that the open-source ASR system WhisperX can transcribe Dutch IDS at a more than sufficient accuracy. We would recommend researchers of IDS compare the performance of multiple off-the-shelf ASR systems in case those are readily available in their language. The accuracy may differ depending on the characteristics of the training data or the data set being transcribed.

In Experiment 1, we found that the difference in recognition performance of Kaldi-NL at transcribing target words in IDS and ADS decreased over time. In Experiment 2, we did not find that Kaldi-NL nor WhisperX performed differently across the two time points when comparing WERs of full sentences. There are two possible explanations for this result. First, utterances could be less affected by the typical acoustic features of IDS compared to target words. Target words (i.e., content words) are typically stressed, while function words are not. In Dutch, stressed syllables are marked by a longer duration and high frequency emphasis [45]. Secondly, previous work has found that when mothers are reading a picture book containing target words to their infants, mothers consistently position these words on exaggerated pitch peaks in utterance-final position [46]. When addressing infants, mothers lengthen the vowels of content words regardless of their position in the utterance [47], while they lengthen the vowels of function words only in utterance-final position [48]. By examining target words (i.e., content words only) rather than utterances (i.e., containing content words and function words) in Experiment 1, the prosodic modifications of IDS may have been more prominent. Since the prosody of IDS becomes more similar to ADS when children grow older, this may have caused a difference in ASR accuracy for target words across the two time points (Experiment 1), which disappeared when analysing full utterances, which are less affected by the prosodic modifications of IDS overall (Experiment 2). An alternative explanation is that Kaldi-NL relies more on the acoustic model compared to WhisperX. That would suggest that Kaldi-NL is more heavily affected by the acoustic differences of IDS, which are more prominent when children are younger. For Kaldi-NL, we found a decrease in WER of 3.7% for speech directed at older children, which is what we would expect if the system relies more heavily on the acoustic model. In contrast, we found an increase of 1.3% in WER for speech directed at older children for WhisperX. If WhisperX relies more on the language model, this could suggest that the performance of WhisperX is more impacted by the low frequency target words that were spoken to older children rather than on the acoustic differences across the two time points.

Future studies should examine whether we can improve the automatic annotation of IDS by applying front-end lowering of mean pitch of the speech recordings (see [49] for the application of this method to children's speech). This could be an efficient, cost-effective solution that can be easily applied by researchers studying different languages—provided a well-trained ASR system in their language exists. This solution, if successful, could be a simple method to create a small but significant improvement in recognition accuracy. Another approach that could be taken in future studies would be to train new language and/or acoustic models on IDS data. For this to work, the IDS data set must be large and general enough to be useful for application on new data sets.

6. Conclusions

In these experiments, we showed that open-source ASR systems can be used for the annotation of Dutch IDS. Although the performance decreased when transcribing IDS compared to ADS, the results are a promising start. Depending on the research goals, automatic transcriptions still need to be corrected by a human annotator. However, this correction process will take less time compared to transcribing the data from scratch. We additionally showed that the choice of ASR system has a large influence on the results. For our Dutch IDS data set, WhisperX performed significantly better than Kaldi-NL. This is the first study that assessed the accuracy of automatic transcriptions of (Dutch) IDS directed at children of different ages generated by different off-the-shelf ASR systems. While there is much room for improvement, the results show that automatic transcriptions provide a promising starting point for researchers who have to transcribe a large amount of speech directed at infants.

Author Contributions: Conceptualization, A.v.d.K., F.A. and R.K.; methodology, A.v.d.K., F.A. and R.K.; formal analysis, A.v.d.K., F.A. and R.K.; data curation, A.v.d.K. and M.H.; writing—original draft preparation, A.v.d.K.; writing—review and editing, F.A., R.K. and M.H.; visualization, A.v.d.K.; supervision, F.A. and R.K.; funding acquisition, R.K. All authors have read and agreed to the published version of the manuscript.

Funding: This work is funded through the Gravitation program of the Dutch Ministry of Education, Culture, and Science and the Netherlands Organization for Scientific Research (NWO grant number 024.001.003) and by Utrecht University's Human-centered Artificial Intelligence focus area (HAI Summer 2021 Small Grant).

Institutional Review Board Statement: There was no ethical committee at the Utrecht Institute of Linguistics (UiL OTS), Utrecht University, when data were collected for this study. This study was approved by UiL OTS and was carried out in accordance with the research guidelines at UiL OTS.

Informed Consent Statement: Informed consent was obtained from all subjects involved in the study.

Data Availability Statement: Anonymised data frames and the *R* markdown script are available online: https://osf.io/jbg9t/, accessed on 23 May 2023.

Acknowledgments: Part of the research reported here was presented at the 2020 International Conference on Multimodal Interaction [50].

Conflicts of Interest: The authors declare no conflict of interest. The funders had no role in the design of the study; in the collection, analyses, or interpretation of data; in the writing of the manuscript; or in the decision to publish the results.

Abbreviations

The following abbreviations are used in this manuscript:

IDS	Infant-directed speech
ADS	Adult-directed speech
ASR	Automatic speech recognition
WER	Word error rate

References

1. Fernald, A.; Simon, T. Expanded intonation contours in mothers' speech to newborns. *Dev. Psychol.* **1984**, *20*, 104–113. [CrossRef]
2. Fernald, A.; Taeschner, T.; Dunn, J.; Papousek, M.; Boysson-Bardies, B.d.; Fukui, I. A cross-language study of prosodic modifications in mothers' and fathers' speech to preverbal infants. *J. Child Lang.* **1989**, *16*, 477–501. [CrossRef]
3. Kuhl, P.K.; Andruski, J.E.; Chistovich, I.A.; Chistovich, L.A.; Kozhevnikova, E.V.; Ryskina, V.L.; Stolyarova, E.I.; Sundberg, U.; Lacerda, F. Cross-language analysis of phonetic units in language addressed to infants. *Science* **1997**, *277*, 684–686. [CrossRef] [PubMed]

4. Soderstrom, M. Beyond babytalk: Re-evaluating the nature and content of speech input to preverbal infants. *Dev. Rev.* **2007**, *27*, 501–532. [CrossRef]
5. Johnson, E.K.; Lahey, M.; Ernestus, M.; Cutler, A. A multimodal corpus of speech to infant and adult listeners. *J. Acoust. Soc. Am.* **2013**, *134*, 534–540. [CrossRef]
6. Han, M.; de Jong, N.H.; Kager, R. Language Specificity of Infant-directed Speech: Speaking Rate and Word Position in Word-learning Contexts. *Lang. Learn. Dev.* **2021**, *17*, 221–240. [CrossRef]
7. Spinelli, M.; Fasolo, M.; Mesman, J. Does prosody make the difference? A meta-analysis on relations between prosodic aspects of infant-directed speech and infant outcomes. *Dev. Rev.* **2017**, *44*, 1–18. [CrossRef]
8. Song, J.Y.; Demuth, K.; Morgan, J. Effects of the acoustic properties of infant-directed speech on infant word recognition. *J. Acoust. Soc. Am.* **2010**, *128*, 389–400. [CrossRef] [PubMed]
9. Zangl, R.; Klarman, L.; Thal, D.; Fernald, A.; Bates, E. Dynamics of Word Comprehension in Infancy: Developments in Timing, Accuracy, and Resistance to Acoustic Degradation. *J. Cogn. Dev.* **2005**, *6*, 179–208. [CrossRef] [PubMed]
10. Singh, L.; Nestor, S.; Parikh, C.; Yull, A. Influences of infant-directed speech on early word recognition. *Infancy* **2009**, *14*, 654–666. [CrossRef]
11. Estes, K.G.; Hurley, K. Infant-directed prosody helps infants map sounds to meanings. *Infancy* **2013**, *18*, 797–824. [CrossRef] [PubMed]
12. Han, M.; de Jong, N.H.; Kager, R. Pitch properties of infant-directed speech specific to word-learning contexts: a cross-linguistic investigation of Mandarin Chinese and Dutch. *J. Child Lang.* **2020**, *47*, 85–111. [CrossRef] [PubMed]
13. Fernald, A. Four-month-old infants prefer to listen to motherese. *Infant Behav. Dev.* **1985**, *8*, 181–195. [CrossRef]
14. Dunst, C.; Gorman, E.; Hamby, D. Preference for infant-directed speech in preverbal young children. *Cent. Early Lit. Learn.* **2012**, *5*, 1–13.
15. Soderstrom, M. ManyBabies1: Infants' preference for infant-directed speech. *J. Acoust. Soc. Am.* **2019**, *145*, 1728–1728. [CrossRef]
16. Cristia, A.; Seidl, A. The hyperarticulation hypothesis of infant-directed speech. *J. Child Lang.* **2014**, *41*, 913–934. [CrossRef]
17. Miyazawa, K.; Shinya, T.; Martin, A.; Kikuchi, H.; Mazuka, R. Vowels in infant-directed speech: More breathy and more variable, but not clearer. *Cognition* **2017**, *166*, 84–93. [CrossRef]
18. Adriaans, F.; Swingley, D. Prosodic exaggeration within infant-directed speech: Consequences for vowel learnability. *J. Acoust. Soc. Am.* **2017**, *141*, 3070–3078. [CrossRef]
19. Kitamura, C.; Thanavishuth, C.; Burnham, D.; Luksaneeyanawin, S. 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 Behav. Dev.* **2001**, *24*, 372–392. [CrossRef]
20. Sjons, J.; Hörberg, T.; Östling, R.; Bjerva, J. Articulation rate in Swedish child-directed speech increases as a function of the age of the child even when surprisal is controlled for INTERSPEECH. *arXiv* **2017**, arXiv:1706.03216.
21. Barras, C.; Geoffrois, E.; Wu, Z.; Liberman, M. Transcriber: Development and use of a tool for assisting speech corpora production. *Speech Commun.* **2001**, *33*, 5–22. [CrossRef]
22. Gaur, Y.; Lasecki, W.S.; Metze, F.; Bigham, J.P. The effects of automatic speech recognition quality on human transcription latency. In Proceedings of the 13th International Web for All Conference, New York, NY, USA, 11–13 April 2016; pp. 1–8. [CrossRef]
23. Burnham, D.; Kalashnikova, M.; Muawiyath, S.; Cassidy, S.; Estival, D. Infant-Directed Speech Research Made Easy: A Database, Some Tools and a Virtual Laboratory. In Proceedings of the Abstract and Paper Presented at the 43rd Experimental Psychology Conference, Melbourne, Australia, 30 March–2 April 2016.
24. Mohamed, A.; Hinton, G.; Penn, G. Understanding how Deep Belief Networks perform acoustic modelling. In Proceedings of the 2012 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Kyoto, Japan, 25–30 March 2012; pp. 4273–4276. [CrossRef]
25. Goldwater, S.; Jurafsky, D.; Manning, C.D. Which words are hard to recognize? Prosodic, lexical, and disfluency factors that increase speech recognition error rates. *Speech Commun.* **2010**, *52*, 181–200. [CrossRef]
26. Kawahara, T.; Nanjo, H.; Shinozaki, T.; Furui, S. Benchmark test for speech recognition using the corpus of spontaneous Japanese. In Proceedings of the ISCA/IEEE Workshop on Spontaneous Speech Processing and Recognition, Tokyo, Japan, 13–16 April 2003.
27. Shinozaki, T.; Hori, C.; Furui, S. Towards automatic transcription of spontaneous presentations. In Proceedings of the Seventh European Conference on Speech Communication and Technology, Aalborg, Denmark, 3–7 September 2001; pp. 491–494.
28. Hirschberg, J.; Litman, D.; Swerts, M. Prosodic and other cues to speech recognition failures. *Speech Commun.* **2004**, *43*, 155–175. [CrossRef]
29. Kirchhoff, K.; Schimmel, S. Statistical properties of infant-directed versus adult-directed speech: insights from speech recognition. *J. Acoust. Soc. Am.* **2005**, *117*, 2238–2246. [CrossRef] [PubMed]
30. Shinozaki, T.; Ostendorf, M.; Atlas, L. Characteristics of speaking style and implications for speech recognition. *J. Acoust. Soc. Am.* **2009**, *126*, 1500–1510. [CrossRef]
31. Han, M. The Role of Prosodic input in Word Learning: A Cross-Linguistic Investigation of Dutch and Mandarin Chinese Infant-Directed Speech. Ph.D Dissertation, Utrecht University, Utrecht, The Netherlands, 2019.
32. Keuleers, E.; Brysbaert, M.; New, B. SUBTLEX-NL: A new measure for Dutch word frequency based on film subtitles. *Behav. Res. Methods* **2010**, *42*, 643–650. [CrossRef]

33. Boersma, P.; Weenink, D. Praat: Doing Phonetics by Computer [Computer Program]. Version 6.1.09. Available online: https://www.fon.hum.uva.nl/praat/ (accessed on 23 May 2023).
34. Yilmaz, E.; Gompel, M. Automatic Transcription of Dutch Speech Recordings [ASR Tool]. 2020. Available online: https://webservices.cls.ru.nl/asr_nl (accessed on 18 March 2020).
35. Oostdijk, N. The Spoken Dutch Corpus. Overview and First Evaluation. In Proceedings of the Second International Conference on Language Resources and Evaluation, Athens, Greece, 31 May–2 June 2000.
36. Peddinti, V.; Wang, Y.; Povey, D.; Khudanpur, S. Low Latency Acoustic Modeling Using Temporal Convolution and LSTMs. *IEEE Signal Process. Lett.* **2018**, *25*, 373–377. [CrossRef]
37. Tejedor-García, C.; van der Molen, B.; van den Heuvel, H.; van Hessen, A.; Pieters, T. Towards an Open-Source Dutch Speech Recognition System for the Healthcare Domain. In Proceedings of the Thirteenth Language Resources and Evaluation Conference, Marseille, France, 20–25 June 2022; pp. 1032–1039.
38. Goutte, C.; Gaussier, E. A probabilistic interpretation of precision, recall and F-score, with implication for evaluation. In *Advances in Information Retrieval*; Losada, D.E., Fernández-Luna, J.M., Eds.; Lecture Notes in Computer Science; Springer: Berlin/Heidelberg, Germany, 2005; pp. 345–359. [CrossRef]
39. Bates, D.; Mächler, M.; Bolker, B.; Walker, S. Fitting linear mixed-effects models using lme4. *J. Stat. Softw.* **2015**, *67*, 1–48. [CrossRef]
40. R Core Team. R: A Language and Environment for Statistical Computing, 2022. Version 4.2.0. Available online: https://www.r-project.org/ (accessed on 23 May 2023).
41. Han, M.; de Jong, N.H.; Kager, R. Relating the prosody of infant-directed speech to children's vocabulary size. *J. Child Lang.* **2023**, 1–17. [CrossRef] [PubMed]
42. Bain, M.; Huh, J.; Han, T.; Zisserman, A. WhisperX: Time-Accurate Speech Transcription of Long-Form Audio. *arXiv* **2023**, arXiv:2303.00747.
43. Radford, A.; Kim, J.W.; Xu, T.; Brockman, G.; McLeavey, C.; Sutskever, I. Robust Speech Recognition via Large-Scale Weak Supervision. *arXiv* **2022**, arXiv:2212.04356.
44. SCTK, the NIST Scoring Toolkit. 2023. Version 2.4.12. Available online: https://github.com/usnistgov/SCTK (accessed on 23 May 2023).
45. Sluijter, A.; van Heuven, V. Acoustic correlates of linguistic stress and accent in Dutch and American English. In Proceedings of the Fourth International Conference on Spoken Language Processing, Philadelphia, PA, USA, 3–6 October 1996; Volume 2, pp. 630–633. [CrossRef]
46. Fernald, A.; Mazzie, C. Prosody and focus in speech to infants and adults. *Dev. Psychol.* **1991**, *27*, 209–221. [CrossRef]
47. Swanson, L.A.; Leonard, L.B.; Gandour, J. Vowel Duration in Mothers' Speech to Young Children. *J. Speech Lang. Hear. Res.* **1992**, *35*, 617–625. [CrossRef] [PubMed]
48. Swanson, L.A.; Leonard, L.B. Duration of function-word vowels in mothers' speech to young children. *J. Speech Hear. Res.* **1994**, *37*, 1394–1405. [CrossRef] [PubMed]
49. Gustafson, J.; Sjölander, K. Voice transformations for improving children's speech recognition in a publicly available dialogue system. In Proceedings of the 7th International Conference on Spoken Language Processing, ISCA, Denver, CO, USA, 16–20 September 2002; pp. 297–300. [CrossRef]
50. van der Klis, A.; Adriaans, F.; Han, M.; Kager, R. Automatic Recognition of Target Words in Infant-Directed Speech. In Proceedings of the Companion Publication of the 2020 International Conference on Multimodal Interaction, Online, 25–29 October 2020; p. 522. [CrossRef]

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.