Automatic Recognition of Target Words in Infant-Directed Speech

Anika van der Klis
Utrecht University
a.vanderklis@uu.nl

Frans Adriaans
Utrecht University
f.w.adriaans@uu.nl

Mengru Han
East China Normal University
mrhan@zhwx.ecnu.edu.cn

René Kager
Utrecht University
r.w.j.kager@uu.nl

ABSTRACT

This study assesses the performance of a state-of-the-art automatic speech recognition (ASR) system at extracting target words in two different speech registers: infant-directed speech (IDS) and adult-directed speech (ADS). We used the Kaldi-NL ASR-service, developed by the Dutch Foundation of Open Speech Technology [5].

IDS is the speech register used by adults when addressing infants. It is characterized by various acoustic and syntactic changes compared to ADS [2]. IDS has also been found to be more variable than ADS [1], [4]. This could pose challenges for an ASR system based on ADS. Since manually transcribing speech recordings is time-consuming, this study aims to assess to what extent an off-the-shelf ASR system trained on ADS can correctly identify target words in IDS. If performance is similar to ADS, then ASR could be used as a research tool. Negative performance constitutes evidence that new tools need to be developed.

Twenty-two Dutch mothers read a picture book to their 18-month-old infant and the experimenter. The picture book was designed to elicit seven disyllabic target words (e.g., walnoot "walnut", kasteel "castle"). The mothers returned six months later and read another picture book eliciting a new set of seven disyllabic target words, mostly of lower frequency (e.g., bamboe "bamboo", jasmijn "jasmine"). The speech recordings were automatically transcribed using the Kaldi-NL ASR-service [5]. We compared the automatic annotations to manual annotations obtained in previous work. This produced the number of "hits" (correctly identified target words), "misses" (missed target words), and "false alarms" (words incorrectly identified as target words). We used these measures to calculate three common accuracy scores: recall, precision, and F-score. Recall represents how complete the extraction of target words was, precision shows the exactness of the recalls, and the F-score is the harmonic mean of recall and precision [3].

At 18 months, recall for target words in IDS is 15.6% lower than for ADS. At 24 months, the difference is only 6.7%. At this age, IDS has already become more similar to ADS, which is reflected in our results. The overall lower scores at 24 months can be explained by the lower frequency of most of these target words. In both registers, precision is 100%. Precision is calculated using false alarms, and there were none. For false alarms to occur, other produced words must be similar to the target words, which is unlikely given the limited contents of the picture books.

The results indicate that, particularly at 18 months, accuracy in IDS is much lower than in ADS. There are differences between IDS and ADS which negatively affect the performance of the existing ASR system. Therefore, new tools need to be developed for the automatic annotation of IDS. Nevertheless, the ASR system can already find more than half of the target words, which is promising.

Table 1: Results of the evaluation procedure

	18 months		24 months	
Register	ADS	IDS	ADS	IDS
Recall	71%	55.4%	60.5%	53.8%
Precision	100%	100%	100%	100%
F-score	83.3%	71.3%	75.4%	70%

CCS CONCEPTS

• **Computing methodologies** → **Speech recognition**; • **Applied computing** → *Arts and humanities.*

KEYWORDS

automatic speech recognition; keyword extraction; infant-directed speech; speech registers

REFERENCES

[1] Alejandrina Cristia and Amanda Seidl. 2014. The hyperarticulation hypothesis of infant-directed speech. *Journal of Child Language* 41, 4 (July 2014), 913–934. https://doi.org/10.1017/S0305000912000669 Publisher: Cambridge University Press.

[2] Anne Fernald, Traute Taeschner, Judy Dunn, Mechthild Papousek, Bénédicte de Boysson-Bardies, and Ikuko Fukui. 1989. A cross-language study of prosodic modifications in mothers' and fathers' speech to preverbal infants. *Journal of Child Language* 16, 3 (Oct. 1989), 477–501. https://doi.org/10.1017/S0305000900010679 Publisher: Cambridge University Press.

[3] Cyril Goutte and Eric Gaussier. 2005. A probabilistic interpretation of precision, recall and F-score, with implication for evaluation. In *Advances in Information Retrieval (Lecture Notes in Computer Science)*, David E. Losada and Juan M. Fernández-Luna (Eds.). Springer, Berlin, Heidelberg, 345–359. https://doi.org/10.1007/978-3-540-31865-1_25

[4] Kouki Miyazawa, Takahito Shinya, Andrew Martin, Hideaki Kikuchi, and Reiko Mazuka. 2017. Vowels in infant-directed speech: More breathy and more variable, but not clearer. *Cognition* 166 (2017), 84–93. https://doi.org/10.1016/j.cognition.2017.05.003

[5] Stichting Open Spraaktechnologie. 2019. https://openspraaktechnologie.org/

ICMI '20 Companion, October 25–29, 2020, Virtual event, Netherlands

ACM ISBN 978-1-4503-8002-7/20/10.
https://doi.org/10.1145/3395035.3425184