Child Speech Recognition in Human-Robot Interaction: Problem Solved?

Ruben Janssens*, Eva Verhelst*, Giulio Antonio Abbo, Qiaoqiao Ren, Maria Jose Pinto Bernal, and Tony Belpaeme

IDLab-AIRO, Ghent University – imec, Ghent, Belgium
{firstname.lastname}@ugent.be

Abstract. Automated Speech Recognition shows superhuman performance for adult English speech on a range of benchmarks, but disappoints when fed children's speech. This has long sat in the way of child-robot interaction. Recent evolutions in data-driven speech recognition, including the availability of Transformer architectures and unprecedented volumes of training data, might mean a breakthrough for child speech recognition and social robot applications aimed at children. We revisit a study on child speech recognition from 2017 and show that indeed performance has increased, with newcomer OpenAI Whisper doing markedly better than leading commercial cloud services. Performance improves even more in highly structured interactions when priming models with specific phrases. While transcription is not perfect yet, the best model recognises 60.3% of sentences correctly barring small grammatical differences, with sub-second transcription time running on a local GPU, showing potential for usable autonomous child-robot speech interactions.

Keywords: Child-Robot Interaction; Automatic Speech Recognition; Verbal Interaction; Interaction Design Recommendations

1 Introduction and Background

Spoken language interaction is for many the holy grail in HCI and HRI. It is built upon a collection of technologies, such as Automated Speech Recognition (ASR), Dialogue Management, and Text-to-Speech, that are chained together to create a system which allows the user to interact or converse with an artificial system using the most natural interface known to humankind. While this processing chain is brittle, the point of entry is Automated Speech Recognition. The ability to automatically transcribe speech utterances has been studied extensively in academic and industrial research. In recent decades, ASR performance has come along in leaps and bounds, with companies claiming "super-human performance" on conversational ASR benchmarks in 2017. On certain benchmarks and for resource-rich languages, speech recognition performance is on par or even better than mean human transcription performance [17]. The popular metric for ASR

* Equal contribution and joint first authors.

performance is Word Error Rate (WER), calculated as the total number of errors —substitutions, insertions, and deletions— divided by the total number of words in the text. WER was typically reported to be below 5% [17]. However, while impressive, these systems' performance degraded catastrophically on speech for which it was not optimised, including atypical voices such as the speech of elderly or young children. This has repercussions for HRI and specifically for applications in which autonomous social robots are expected to interact with non-typical users, such as robots for elder care or robots for education [2, 4, 16].

In 2017, Kennedy *et al.* [8] published a widely cited study showing that then state-of-the-art ASR could not reliably transcribe the speech of 5-year-old English speakers. The speech ranged from constrained utterances —such as counting from 1 to 10— to unconstrained telling of a story from a picture book. The ASR performance was evaluated for four different engines, but the results were nothing but disappointing. While WER for adult speech was below 5%, most engines could not correctly transcribe a single child utterance. Only Google's ASR did marginally better, recognising 11.8% of constrained child speech and about 6% of spontaneous child speech. Still, only correctly being able to transcribe 1 utterance out of 10 is a recipe for interaction disaster, and the authors of the study then recommended against relying on ASR for child-robot interaction.

Forward 6 years. Artificial intelligence has been revolutionised by the Transformer architecture, not only resulting in a sea of change in the performance of generative language models but also in the performance of ASR [11]. In September 2022, OpenAI released Whisper, an ASR engine built using an encoder-decoder Transformer architecture trained on an unprecedented 680,000 hours of labelled audio data [13]. While the specifics of Whisper's training regimen and its training data are proprietary to OpenAI, the inference model is released as public open-source software. Whisper's performance on average is better than competing solutions, but was found to still be subpar to solutions that have been specifically trained or fine-tuned on specific datasets, such as LibriSpeech [13].

Next to the publicly available Whisper models, there are several cloud-based solutions. In this area large players —Amazon, Google, Microsoft and Tencent— compete with smaller, sometimes specialised vendors, but all offer convenient online API services that are easily integrated within code.

Given the availability of new architectures trained on larger and more diverse corpora, the time is opportune to revisit the results from Kennedy *et al.* [8] and evaluate whether state-of-the-art ASR can now handle child speech. Recent work on child speech recognition has focused mainly on finetuning open source ASR models that were originally trained on adult speech on additional child speech data (e.g., [5], [3], [6], [1]), resulting in a decreased WER of up to 5% [1], but this increase in performance may not be generalizable to other datasets [5]. Additionally, as little child speech data is available, approaches like data augmentation [7], combining datasets of different languages [14] or parameter-efficient finetuning [12] are used. We choose to evaluate readily available ASR systems, including commercial models, as our focus is the out-of-the-box applicability of these systems in child-robot interaction.

We decided to compare OpenAI's Whisper, as it is open-source and exemplifies the new direction in data-driven ASR, and two commercial cloud-based solutions, opting for Microsoft and Google's systems. We are first and foremost interested in transcription accuracy, but for our aim of integrating child speech recognition into a real-time interactive HRI scenario, we also explore how responsive different systems are. We also investigate whether performance can be improved by priming the systems to expect specific words or phrases.

2 Methodology

To evaluate the ASR engines, we use the data from Kennedy *et al.* [9] which contains audio recordings of 11 young children (age M=4.9 years old) recorded at an English primary school. The recordings consist of spontaneous speech (retelling a picture book) and speech in which children count from 1 to 10 or repeat short sentences spoken by an adult (such as "the horse is in the stable"). Each sample is recorded from 3 sources: a studio-grade microphone (Rode NT1-A) placed above the robot, a portable microphone (Zoom H1) placed just in front of the robot, and the two front microphones of the Aldebaran NAO robot. All recordings were manually transcribed and this is used as ground truth.

We evaluated three ASR engines: Microsoft Azure Speech to Text, Google Cloud Speech-to-text, and OpenAI's Whisper. The Azure and Google models were used through a cloud API. Whisper exists in different model sizes: tiny (39M parameters), base, small, medium, and large (1550M parameters), with three versions of the large model. All seven of these models are compared in this study: we expect the smaller models to run faster but have lower accuracy. We used the `faster-whisper`[1] reimplementation of the Whisper models, which claims a transcription time of up to four times faster than OpenAI's original Whisper implementation. The Whisper models were run locally on an NVIDIA GeForce GTX 1080 Ti with 11 GB of VRAM. We also ran them on only a CPU, to assess the necessity of a dedicated GPU for these models. We configured the models to expect English language speech, as preliminary testing revealed that without this option Whisper large v3 correctly detected English in only 84% of spontaneous speech samples. All transcriptions were performed in 2024.

The performance of the models is compared using four different metrics for transcription accuracy, to ensure comparability with the results reported in [8]:

- **Levenshtein distance**: the minimum amount of insertions, deletions and substitutions required to change one sequence into the other, at letter level, divided by the amount of letters in the ground truth. A score of 0 means perfect recognition, a score of 1 could reflect a recognised sequence of the same length but with no letter in the right position. There is no upper bound.
- **WER**, for comparability with other ASR research.
- **Absolute accuracy**: the fraction of samples that are completely correctly recognised.

[1] github.com/SYSTRAN/faster-whisper

– **Relaxed accuracy**: the fraction of samples that are correctly recognised, also counting as accurate those with small grammatical differences that do not impact the meaning of the utterance, following the same rules as in [8].

To estimate the possibility of real-time interactions, we explore the responsiveness of the different systems by reporting their transcription time. For all Whisper models, the transcription time is the time it takes for the model to return a result, which varies due to the model size as well as the hardware on which it runs. As the Azure and Google systems are cloud-based, their transcription time also includes the transmission time of the audio file and the result.

In all analyses, unless otherwise stated, we use only the studio microphone recordings, and only the recordings of the sentences that the children repeat from the adult ($n = 50$) and of the spontaneous utterances (split into sentences, $n = 222$), because preliminary analysis showed that utterances consisting of a single number are often too short for the engines to detect any speech.

3 Results

We will first compare the models' transcription accuracy, also investigating the impact of the length of the utterance, and improving performance by using model priming techniques. Then, we report the models' responsiveness, the impact of the microphone used, and finally a reflection on the power consumption of the models.

3.1 Transcription accuracy

First of all, we compare the performance of Google, Azure and the best Whisper model (large v3) with the four engines reported in the 2017 paper. These results are shown in Figure 1. They show that the Google speech recognition did not improve compared to 2017, but the performance of both the Azure model, and the Whisper model are better than all models tested in the 2017 paper, with Whisper performing best of all.

We report the transcription accuracy using multiple metrics in Table 1. All four metrics show the same ordering between the models' performances.

While WER was not reported in the 2017 paper, it allows us to compare with performance on adult speech. Microsoft achieved a WER of 5% in 2017 on part of the Switchboard dataset, but another part of this dataset was used for training the model [17]. In 2023, Whisper achieved a WER of 13.8% on Switchboard without being specifically trained for this dataset. These results suggest that Whisper's generalisability across datasets extends to child speech, although there is still a gap with adult speech.

The absolute and relaxed accuracy give an impression of the usability of the models. While this is not yet an ideal performance level, the relaxed accuracy shows that Whisper is already rather usable. Some small mistakes could still be handled by dialogue management software, such as large language models,

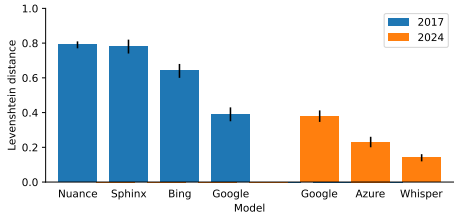


Fig. 1: Performance of ASR engines in 2017 and 2024, calculated as mean normalised Levenshtein distance between ground truth and transcription (lower is better).

Table 1: Transcription accuracy for ASR engines in 2017 and 2024, for repeated sentences and spontaneous utterances split into sentences ($n = 272$).

Metric	**Google (2017)**	**Google (2024)**	**Azure**	**Whisper**
Levenshtein distance (↓ better)	0.38	0.38	0.23	**0.14**
WER (↓ better)	-	49.0%	30.3%	**21.3%**
Absolute accuracy (↑ better)	7.5%	9.6%	23.5%	**36.8%**
Relaxed accuracy (↑ better)	20.3%	14.7%	43.0%	**60.3%**

even though they do not count as accurate in the relaxed accuracy criteria. For example, in the sentence "the dog is in front of the horse", Whisper and Azure replace "in" with "the", but this transcription could still be used to further the conversation, even though it is counted as inaccurate in these metrics.

Figure 2a shows a more detailed comparison between the different model sizes of Whisper and the Azure and Google services. As expected, the large Whisper models perform best, with Whisper large v3 performing best of all.

The Kruskal-Wallis test reveals significant differences between the Levenshtein distance for the tested models ($p < 0.001$), and post-hoc Dunn tests with Bonferroni correction do not show significant differences between Whisper large v3 and Whisper small, but do show a significant difference between, among others, all large Whisper models and Azure ($p < 0.005$), and between Azure and Google ($p < 0.001$).

3.2 Utterance length

The previous section reported performance on the samples containing sentences that the children repeated from an adult, and spontaneous utterances that were split into sentences. However, the dataset also contains samples containing a number that was repeated from the adult, as well as full-length spontaneous utterances. These samples are respectively much shorter (mean length 0.98s, s.d. 0.43s) and much longer (108.91s, s.d. 69.6s) than the repeated sentences (2.85s, s.d. 0.61s) and spontaneous speech cut into sentences (4.20s, s.d. 2.09s).

As shown in Table 2, the ASR systems perform considerably worse on these shorter and longer samples. Only half of the repeated numbers are correctly recognized. Looking at the mistakes the models made shows that the samples

containing a number are often too short to detect any speech. On the other hand, the samples containing the full spontaneous speech fragments contain pauses between sentences, which sometimes lead to early stopping of the transcription.

Table 2: Levenshtein distance for different utterance types (accuracy in brackets).

	Google	**Azure**	**Whisper**
Repeated sentences	0.26	0.16	0.13
Spontaneous sentences	0.41	0.25	0.14
Repeated numbers	0.80 (17%)	0.38 (52%)	0.50 (48%)
Full spontaneous samples	0.43	0.89	0.55

3.3 Model priming

All three ASR systems we evaluated offer functionalities to prime the models to expect certain words or phrases. In a setting as in this dataset, where the context of the conversation is known, this could improve performance.

For the repeated numbers, we primed the Google and Azure ASR systems to expect a number—specifically between one and ten for Azure. Whisper works differently: it can be provided with a prompt, which acts as conversational context, which the model will try to complete with the transcript. The prompt *"I choose number"* worked best. As shown in Table 3, model priming strongly improves performance for these samples. Interestingly, the Whisper medium model performs better than large v3, reaching an average Levenshtein distance of 0.17 and an absolute accuracy of 88.0%. Table 3 reports the results of large v3.

For the repeated sentences, we compare three priming approaches: providing the five full sentences that the children said, making this a multiple-choice task, providing a template grammar, as also described in [8], and providing the separate words and subphrases that are part of the template. For Whisper, the template is provided as *"Choose the words that were said by this child: "The <dog/fish/horse> is <in/next to/in front of/behind/on top of> the <pond/shed/car/stable/horse>""*. Instead of providing separate words to Whisper, the prompt *"Where is the animal?"* is used. Table 3 also shows improved performance. As expected, the more structured the priming becomes, the better the performance: Whisper achieves almost perfect recognition in the multiple-choice setting.

For the spontaneous speech samples, Google and Azure were primed with the animals that were part of the template, while Whisper was prompted with *"Today we'll read 'Frog, where are you?'. This book chronicles the humorous adventures that befall a young boy as he and his dog search for the frog that escaped from a jar in his room. Let's begin.*" Here, the performance gain is more limited. For Whisper, this translates into a WER for the spontaneous sentences of 21.9% with the prompt compared to 23.1% without.

3.4 Responsiveness

In Figure 2b, the average transcription times for short sentences (spontaneous speech and repeat sentences) are shown for Google, Azure, all of the Whisper

Table 3: Levenshtein distance for primed models (accuracy between brackets).

Utterance type	Prompt	Google	Azure	Whisper
Repeated numbers	None	0.80 (17%)	0.38 (52%)	0.77 (40%)
	Number	0.66 (33%)	0.32 (64%)	0.53 (72%)
Repeated sentences	None	0.26	0.16	0.14
	Separate words	0.22	0.13	0.09
	Template	0.18	unsupported	0.07
	Multiple choice	0.20	0.09	0.01
Spontaneous sentences	None	0.41	0.25	0.15
	Animals	0.39	0.24	0.13
Full spontaneous samples	None	0.43	0.89	0.20
	Animals	0.42	0.89	0.19

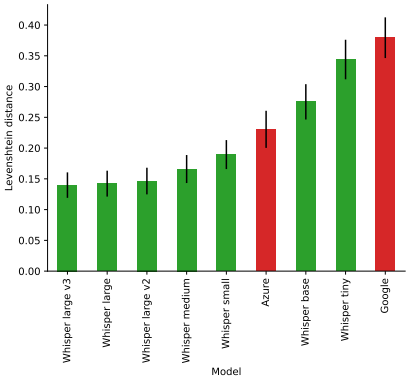
models on GPU and the tiny, base and small Whisper models on CPU. The average transcription time for the Whisper medium and large models on CPU are respectively 17.5s and 30.5s and were left out of the graph for readability. We marked the 1000ms line on the figure because, even though the mean response time in human conversation is 200ms, for spoken dialogue systems a delay of between 700 and 1000ms is deemed acceptable [15]. This data shows that using a model locally on a GPU, instead of CPU or using an API, can greatly improve the responsiveness, even reaching an acceptable level for spoken dialogue.

The Kruskal-Wallis test shows significant differences between the transcription time of the tested models ($p < 0.001$), and post-hoc Dunn tests with Bonferroni correction show significant differences between all pairs of models, except for between Whisper tiny, base, and small, between Whisper large v3 and medium, and between Whisper large, large v2, and Azure.

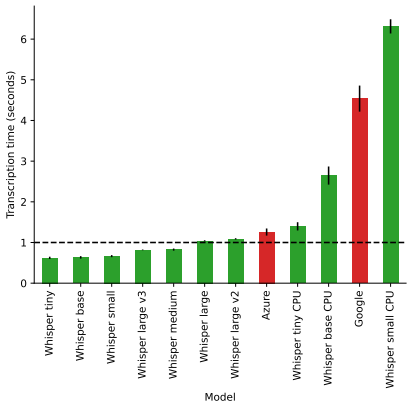
Figure 2c shows the relation between the models' average transcription time with their accuracy using the Levenshtein distance. To choose which model to use, both responsiveness and performance should be taken into account. Lower results are preferred for both, so models in the lower left corner of the scatter plot are ideal. As apparent in the figure, there is a trade-off between transcription time and transcription performance, so the choice should be made based on the specific application.

3.5 Microphone

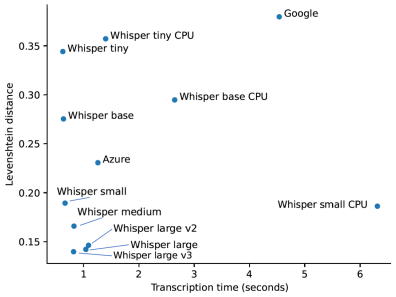
In Figure 2d, we compare the transcription accuracy between the three microphones used to record the samples, using the average performance of Google, Azure and Whisper large v3 on each sample. When comparing the results of the internal Nao microphone with the portable and studio microphone, the Kruskal-Wallis test shows a significant difference between the groups, and Dunn's test with Bonferroni corrections as post-hoc analysis shows a significant difference between the Nao and portable microphone ($p < 0.001$) and the Nao and studio microphone ($p < 0.01$). There is no significant difference between the studio and portable microphone ($p = 0.399$). In conclusion, the worst results are obtained when the microphone in the Nao robot is used, as there is a lot of added noise



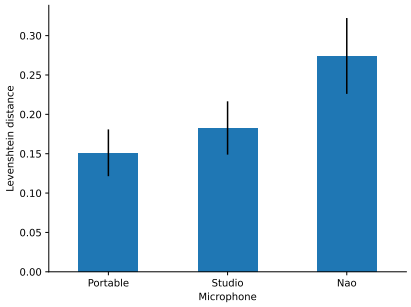
(a) Performance of all current ASR engines (Whisper model versions in green, Google and Azure in red) calculated as Levenshtein distance between ground truth and transcription (lower is better).



(b) Average transcription time for a sentence. Whisper models (in green) were run on GPU unless CPU is specified. Dashed line shows maximum acceptable delay of 1000ms.



(c) Transcription time vs. accuracy (lower is better). Whisper models were run on GPU unless CPU is specified. Ideal ASR systems would be in the lower left corner.



(d) Average performance of Azure, Google and best Whisper model when using different microphone types, calculated as Levenshtein distance (lower is better). Best results are obtained when using a microphone external from the robot.

Fig. 2: Graphs comparing the performance and transcription time of all current ASR engines and the impact of the microphone type on these metrics, for repeated sentences and spontaneous utterances split into sentences ($n = 272$).

due to the closeness to the robot's motor and ventilation, but no difference is found between both external microphones.

3.6 Energy consumption

As concerns have been raised over the energy consumption and consequently carbon emissions of state-of-the-art machine learning [10], we think it is valuable to consider these for ASR systems. We were only able to measure Whisper's consumption, as the other models were used via API instead of on our machines. Per hour of transcribed data, the largest and best-performing model (large v3) consumed $32.3Wh$ and produced $7.7gCO_2eq$ in our setup. Note that this energy consumption refers to the transcription of many samples back-to-back, while real-time interactive systems would not be able to work as efficiently.

4 Conclusion

Based on our evaluation, we can make the following recommendations, updating or overriding those made in [8]:

Recognition performance. The recognition performance has improved dramatically for state-of-the-art ASR, with the best models of 2024 showing over 60% fewer transcription errors than in 2017. However, performance falters when input samples become much shorter or longer than a single sentence. Model priming is able to improve performance, especially in very structured settings, e.g. when a number of fixed options are expected. Adult-like recognition is not available yet, but the semantic content of children's speech is now sufficiently transcribed to offer potential for robust spoken interaction. Small errors can often be caught by other components of the dialogue management system —such as large language models— resulting in a higher task accuracy than the reported accuracy of ASR systems. However, future research should investigate performance in other languages than English.

Responsiveness. The responsiveness of locally hosted models (in our case OpenAI's Whisper) is significantly better than that of cloud-based solutions, with sub-second results for some models, when running them on a GPU. The network overhead and shared services of using cloud-based solutions are not optimal for real-time spoken interaction, and local models even outperform the cloud-based solutions in accuracy.

Impact of microphone. Using an external microphone, as opposed to a microphone embedded in the robot, leads to a significantly improved recognition performance. Performance improves regardless of the quality of the microphone, as the robot's noise has a stronger effect on the speech recognition than the choice of microphone.

Acknowledgments. This research received funding from imec (Smart Education), the Flemish Government (AI Research Program) and the Horizon Europe VALAWAI project (grant agreement number 101070930). We are indebted to the authors of [9] and [8] for making the recordings and transcriptions available.

References

1. Attia, A.A., Liu, J., Ai, W., Demszky, D., Espy-Wilson, C.: Kid-whisper: Towards bridging the performance gap in automatic speech recognition for children vs. adults. arXiv preprint arXiv:2309.07927 (2023)
2. Belpaeme, T., Kennedy, J., Ramachandran, A., Scassellati, B., Tanaka, F.: Social robots for education: A review. Science robotics **3**(21), eaat5954 (2018)
3. Booth, E., Carns, J., Kennington, C., Rafla, N.: Evaluating and improving child-directed automatic speech recognition. In: Proceedings of the Twelfth Language Resources and Evaluation Conference. pp. 6340–6345 (2020)
4. Cifuentes, C.A., Pinto, M.J., Céspedes, N., Múnera, M.: Social robots in therapy and care. Current Robotics Reports **1**, 59–74 (2020)
5. Jain, R., Barcovschi, A., Yiwere, M., Corcoran, P., Cucu, H.: Adaptation of whisper models to child speech recognition. arXiv preprint arXiv:2307.13008 (2023)
6. Jain, R., Barcovschi, A., Yiwere, M.Y., Bigioi, D., Corcoran, P., Cucu, H.: A wav2vec2-based experimental study on self-supervised learning methods to improve child speech recognition. IEEE Access **11**, 46938–46948 (2023)
7. Kathania, H.K., Kadyan, V., Kadiri, S.R., Kurimo, M.: Data augmentation using spectral warping for low resource children asr. Journal of Signal Processing Systems **94**(12), 1507–1513 (2022)
8. Kennedy, J., Lemaignan, S., Montassier, C., Lavalade, P., Irfan, B., Papadopoulos, F., Senft, E., Belpaeme, T.: Child speech recognition in human-robot interaction: evaluations and recommendations. In: Proceedings of the 2017 ACM/IEEE international conference on human-robot interaction. pp. 82–90 (2017)
9. Kennedy, J., Lemaignan, S., Montassier, C., Lavalade, P., Irfan, B., Papadopoulos, F., Senft, E., Belpaeme, T.: Children speech recording (English, spontaneous speech + pre-defined sentences) (Dec 2016). https://doi.org/10.5281/zenodo.200495, https://doi.org/10.5281/zenodo.200495
10. Lacoste, A., Luccioni, A., Schmidt, V., Dandres, T.: Quantifying the carbon emissions of machine learning. arXiv preprint arXiv:1910.09700 (2019)
11. Latif, S., Zaidi, A., Cuayahuitl, H., Shamshad, F., Shoukat, M., Qadir, J.: Transformers in speech processing: A survey. arXiv preprint arXiv:2303.11607 (2023)
12. Liu, W., Qin, Y., Peng, Z., Lee, T.: Sparsely shared lora on whisper for child speech recognition. In: ICASSP 2024-2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). pp. 11751–11755. IEEE (2024)
13. Radford, A., Kim, J.W., Xu, T., Brockman, G., McLeavey, C., Sutskever, I.: Robust speech recognition via large-scale weak supervision. In: International Conference on Machine Learning. pp. 28492–28518. PMLR (2023)
14. Rolland, T., Abad, A., Cucchiarini, C., Strik, H.: Multilingual transfer learning for children automatic speech recognition (2022)
15. Skantze, G.: Turn-taking in conversational systems and human-robot interaction: a review. Computer Speech & Language **67**, 101178 (2021)
16. Verhelst, E., Janssens, R., Demeester, T., Belpaeme, T.: Adaptive second language tutoring using generative ai and a social robot. In: Companion of the 2024 ACM/IEEE International Conference on Human-Robot Interaction. pp. 1080–1084 (2024)
17. Xiong, W., Wu, L., Alleva, F., Droppo, J., Huang, X., Stolcke, A.: The microsoft 2017 conversational speech recognition system. In: 2018 IEEE international conference on acoustics, speech and signal processing (ICASSP). pp. 5934–5938. IEEE (2018)

Child Speech Recognition in Human-Robot Interaction: Problem Solved?

| Conference paper | First Online: 25 March 2025

| pp 476–486 | Cite this conference paper

Social Robotics

(ICSR + AI 2024)

Ruben Janssens, Eva Verhelst, Giulio Antonio Abbo, Qiaoqiao Ren, Maria Jose Pinto Bernal & Tony Belpaeme

Part of the book series: Lecture Notes in Computer Science ((LNAI,volume 15562))

Included in the following conference series:
International Conference on Social Robotics

Abstract

Automated Speech Recognition shows superhuman performance for adult English speech on a range of benchmarks, but disappoints when fed children's speech. This has long sat in the way of child-robot interaction. Recent evolutions in data-driven speech recognition, including the availability of Transformer architectures and unprecedented volumes of training data, might mean a breakthrough for child speech recognition and social robot applications aimed at children. We revisit a study on child speech recognition from 2017 and show that indeed performance has increased, with newcomer OpenAI Whisper doing markedly better than leading commercial cloud services. Performance improves even more in highly structured interactions when priming models with specific phrases. While transcription is not perfect yet, the best model recognises 60.3% of sentences correctly barring small grammatical differences, with sub-second transcription time running on a local GPU, showing potential for usable autonomous child-robot speech interactions.

R. Janssens and E. Verhelst—Equal contribution and joint first authors.

Subscribe now →

Buy Now

Chapter

EUR 29.95
Price includes VAT (Belgium)

Available as PDF
Read on any device
Instant download
Own it forever

Buy Chapter

eBook

EUR 60.98
Price includes VAT (Belgium)

Available as EPUB and PDF
Read on any device
Instant download
Own it forever

Buy eBook

Softcover Book

EUR 78.43
Price includes VAT (Belgium)

Compact, lightweight edition
Dispatched in 3 to 5 business days
Free shipping worldwide - see info

Buy Softcover Book

Notes

1. github.com/SYSTRAN/faster-whisper.

References

1. Attia, A.A., Liu, J., Ai, W., Demszky, D., Espy-Wilson, C.: Kid-whisper: towards bridging the performance gap in automatic speech recognition for children vs. adults. arXiv preprint arXiv:2309.07927 (2023)

2. Belpaeme, T., Kennedy, J., Ramachandran, A., Scassellati, B., Tanaka, F.: Social robots for education: a review. Science robotics 3(21), eaat5954 (2018)

Google Scholar

3. Booth, E., Carns, J., Kennington, C., Rafla, N.: Evaluating and improving child-directed automatic speech recognition. In: Proceedings of the Twelfth Language Resources and Evaluation Conference, pp. 6340–6345 (2020)

Google Scholar

4. Cifuentes, C.A., Pinto, M.J., Céspedes, N., Múnera, M.: Social robots in therapy and care. Curr. Robot. Rep. **1**, 59–74 (2020)

MATH Google Scholar

5. Jain, R., Barcovschi, A., Yiwere, M., Corcoran, P., Cucu, H.: Adaptation of whisper models to child speech recognition. arXiv preprint arXiv:2307.13008 (2023)

6. Jain, R., Barcovschi, A., Yiwere, M.Y., Bigioi, D., Corcoran, P., Cucu, H.: A wav2vec2-based experimental study on self-supervised learning methods to improve child speech recognition. IEEE Access **11**, 46938–46948 (2023)

Google Scholar

7. Kathania, H.K., Kadyan, V., Kadiri, S.R., Kurimo, M.: Data augmentation using spectral warping for low resource children ASR. J. Signal Process. Syst. **94**(12), 1507–1513 (2022)

MATH Google Scholar

8. Kennedy, J., et al.: Child speech recognition in human-robot interaction: evaluations and recommendations. In: Proceedings of the 2017 ACM/IEEE International Conference on Human-Robot Interaction, pp. 82–90 (2017)

Google Scholar

9. Kennedy, J., et al.: Children speech recording (English, spontaneous speech + pre-defined sentences) (2016). https://doi.org/10.5281/zenodo.200495, https://doi.org/10.5281/zenodo.200495

10. Lacoste, A., Luccioni, A., Schmidt, V., Dandres, T.: Quantifying the carbon emissions of machine learning. arXiv preprint arXiv:1910.09700 (2019)

11. Latif, S., Zaidi, A., Cuayahuitl, H., Shamshad, F., Shoukat, M., Qadir, J.: Transformers in speech processing: a survey. arXiv preprint arXiv:2303.11607 (2023)

12. Liu, W., Qin, Y., Peng, Z., Lee, T.: Sparsely shared Lora on whisper for child speech recognition. In: ICASSP 2024-2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pp. 11751–11755. IEEE (2024)

Google Scholar

13. Radford, A., Kim, J.W., Xu, T., Brockman, G., McLeavey, C., Sutskever, I.: Robust speech recognition via large-scale weak supervision. In: International Conference on Machine Learning, pp. 28492–28518. PMLR (2023)

Google Scholar

14. Rolland, T., Abad, A., Cucchiarini, C., Strik, H.: Multilingual transfer learning for children automatic speech recognition (2022)

Google Scholar

15. Skantze, G.: Turn-taking in conversational systems and human-robot interaction: a review. Comput. Speech Langu. **67**, 101178 (2021)

MATH Google Scholar

16. Verhelst, E., Janssens, R., Demeester, T., Belpaeme, T.: Adaptive second language tutoring using generative AI and a social robot. In: Companion of the 2024 ACM/IEEE International Conference on Human-Robot Interaction, pp. 1080–1084 (2024)

Google Scholar

17. Xiong, W., Wu, L., Alleva, F., Droppo, J., Huang, X., Stolcke, A.: The microsoft 2017 conversational speech recognition system. In: 2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pp. 5934–5938. IEEE (2018)

Google Scholar

Acknowledgments

This research received funding from imec (Smart Education), the Flemish Government (AI Research Program) and the Horizon Europe VALAWAI project (grant agreement number 101070930). We are indebted to the authors of [9] and [8] for making the recordings and transcriptions available.

Author information

Authors and Affiliations

IDLab-AIRO, Ghent University – imec, Ghent, Belgium
Ruben Janssens, Eva Verhelst, Giulio Antonio Abbo, Qiaoqiao Ren, Maria Jose Pinto Bernal & Tony Belpaeme

Corresponding author

Correspondence to Ruben Janssens .

Editor information

Editors and Affiliations

University of Southern Denmark, Odense, Denmark
Oskar Palinko

University of Southern Denmark, Odense, Denmark

Leon Bodenhagen

Qatar University, Doha, Qatar
John-John Cabibihan

University of Southern Denmark, Sønderborg, Denmark
Kerstin Fischer

Indiana University, Bloomington, IN, USA
Selma Šabanović

Uppsala University, Uppsala, Sweden
Katie Winkle

IIT Mandi, Mandi, Himachal Pradesh, India
Laxmidhar Behera

National University of Singapore, Singapore, Germany
Shuzhi Sam Ge

Aalborg University, Aalborg, Denmark
Dimitrios Chrysostomou

Qingdao University, Qingdao, China
Wanyue Jiang

The University of Alabama, Tuscaloosa, AL, USA
Hongsheng He

Rights and permissions

Copyright information

About this paper

Cite this paper

Janssens, R., Verhelst, E., Abbo, G.A., Ren, Q., Bernal, M.J.P., Belpaeme, T. (2025). Child Speech Recognition in Human-Robot Interaction: Problem Solved?. In: Palinko, O., *et al.* Social Robotics. ICSR + AI 2024. Lecture Notes in Computer Science(), vol 15562. Springer, Singapore. https://doi.org/10.1007/978-981-96-3519-1_43

.RIS .ENW .BIB

DOI
https://doi.org/10.1007/978-981-96-3519-1_43

Published
25 March 2025

Publisher Name
Springer, Singapore

Print ISBN
978-981-96-3518-4

Online ISBN
978-981-96-3519-1

eBook Packages
Computer Science
Computer Science (R0)

Publish with us

Policies and ethics