Large Language Models Cover for Speech Recognition Mistakes: Evaluating Conversational AI for Second Language Learners

Eva Verhelst
IDLab - AIRO
Ghent University - imec
Ghent, Belgium
eva.verhelst@ugent.be

Tony Belpaeme
IDLab - AIRO
Ghent University - imec
Ghent, Belgium
tony.belpaeme@ugent.be

***Abstract*—Automatic Speech Recognition (ASR) technology has been reported to reach near-human performance in recent years, yet it continues to struggle with atypical speakers, particularly second language learners. This limitation has hindered progress in leveraging social robots for second language education, a field with significant promise. Recent advancements in Large Language Models (LLMs), which demonstrate capabilities in context understanding, common sense reasoning, and pragmatics, offer a potential solution by compensating for transcription errors introduced by ASR. This study examines whether ASR combined with an LLM can produce flowing conversation. Particularly, we look at its application in learning French as a second language by Dutch-speaking students. Through task-based interactions, where successful task completion depends on the accurate interpretation of user speech, the study evaluates the impact of LLMs on conversational outcomes. Results confirm that the performance of ASR degrades significantly for both speakers with limited proficiency and a non-English language. Nonetheless, LLMs demonstrate the ability to interpret context and sustain meaningful conversations despite suboptimal ASR outputs, highlighting a promising path forward for the integration of these technologies in second-language education.**

***Index Terms*—Speech Recognition, Large Language Models, L2 Speakers, Pragmatics**

I. Introduction and Related Work

The first step in achieving fluent spoken interaction between humans and robots – the holy grail of human-robot interaction (HRI) – is to accurately understand the user's speech. While automatic speech recognition (ASR) has historically posed a challenge in HRI, some datasets have demonstrated ASR performance comparable to human transcriptions [1]. However, when moving beyond controlled conditions to the speech patterns of atypical populations, this performance often declines significantly. The amount of resources available for the language and the specifics of the speaker play a substantial role in this.

The use of social robots in language education has been explored extensively, showing it to be a promising field [2]–[5]. ASR is a crucial component in the processing chain, but unfortunately the combination of non-fluent competence and often young speakers, results in suboptimal ASR outcomes, hampering the overall performance of robot language tutors. Studies indicate that modern speech recognition systems can only understand approximately 60% of spoken utterances from young children [6]. Models specifically fine-tuned on child speech yield significant improvements to those only trained on adult speech, but this often does not generalize to other datasets [7]–[9]. Different accents, both for fluent and non-fluent speakers, also strongly influence ASR performance. Graham and Roll compared ASR performance on English speakers with a variety of accents, and found that American accents result in better recognition than other English accents and that second language (L2) speakers' accents further exacerbate recognition [10]. Furthermore, ASR performance tends to decrease when recognising non-English languages [11]. In [12], the ASR performance on Dutch read and conversational speech of first (L1) and second language speakers is compared. As expected, ASR performance is much better for read speech and for L1 speakers. Still, of the results presented there, only the word error rate (WER) for read L1 speech comes close to the human performance level. This suboptimal result might show the extent of the impact of a lower resource language.

This paper examines ASR performance on Flemish students, with Dutch as a first language, speaking French as a second language. Prior research has highlighted that OpenAI's Whisper ASR system outperforms its commercially available competitors for atypical speakers [6], so Whisper will be used for this analysis. Whisper's performance on French is already notably lower compared to English [11], and, given the aforementioned related literature, it is anticipated that performance will be even lower when processing non-fluent speech.

Due to their impressive improvements in the previous years, conversational systems for social robots often include Large Language Models (LLMs). LLMs are known for having a state-of-the-art grasp on semantics, and research has shown that at scale, they might become capable of pragmatics: using context to improve practical language understanding in

This research is funded by imec Smart Education, the Research Foundation Flanders (FWO Vlaanderen, 1S50425N) and the Flanders AI Research 2 initiative.

communication [13]. In humans, our understanding of speech is heavily influenced by our expectations and the context in which the speech takes place. Using this idea to improve speech recognition has been attempted as long as decades ago [14]. Therefore, this paper studies whether the intrinsic pragmatic capabilities in LLMs can improve the practical success of ASR systems in a conversation when they are used in series. Practically, this will be evaluated using the measure of task completion: was the ASR output together with the LLM's pragmatic capabilities enough to formulate an appropriate response? The task used in this paper was designed to allow for an objective evaluation. As the speech used in this paper is that of L2 speakers, we hypothesise that the ASR results will be worse than previously found human-quality results and that the pragmatic capabilities of the LLM will allow for improved task completion.

This paper attempts to answer two research questions and formulates two respective hypotheses.

RQ1 Is ASR performance adequate for HRI in a language learning context where Dutch L1 speakers learn French?

H1 ASR performance of Dutch L1 speakers speaking French will be considerably worse than the superhuman performance seen in benchmarks of fluent English speakers.

RQ2 Is task completion fully dependent on ASR performance, or can wrong ASR transcriptions be repaired by an LLM?

H2 The LLM will cover for enough of the ASR system's mistakes to allow for successful conversations even with inadequate ASR transcriptions.

II. Data Collection

This section describes what data was collected to answer the research questions. This includes a description of the user study, an overview of the interaction users had with the robot and a summary of the data that was collected.

A. User Study

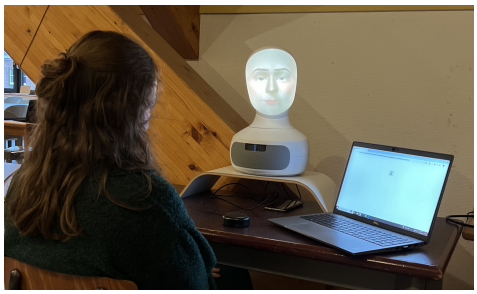

Fig. 1: Study set-up consisting of a Furhat and a laptop, on which the story was displayed.

As the first research question focuses on the ASR performance of second language learners, more specifically Dutch (Flemish) speakers learning French, the participants were chosen to be secondary school students that follow a French course. To ensure sufficient language skills, allowing for conversation, students in their fourth year of secondary school were chosen, which corresponds to their sixth year of French classes. One secondary school class participated in the data collection exercise ($n = 17$, mean age = 14.8, of which 15 identified as female). Recruitment happened through the school, and informed consent was obtained from the participants, their parents or guardians, and the school. The research was conducted according to the ethical rules presented in the General Ethics Protocol of the faculty of Engineering and Architecture of Ghent University. The participants were briefed about the interaction, including a short demonstration. Figure 1 shows the set-up of the user study.

B. Interaction

The data collection process involved an interactive storytelling activity facilitated by a Furhat robot. The student is prompted to enter a topic for the story via the user interface. Based on this topic, the robot initiated a narrative generated by a large language model (LLM), which was designed to structure the story into seven conversational turns. At the end of each turn, the robot ended with a question for the student concerning the progression of the narrative (e.g., "Where does she go next?", "What should she do with this treasure?"). The participant's response directly influenced the subsequent direction of the story. An example of an interaction is shown in Figure 2. Task completion was evaluated based on the extent to which the participant's input was successfully incorporated into the narrative, thereby serving as the metric for addressing RQ2.

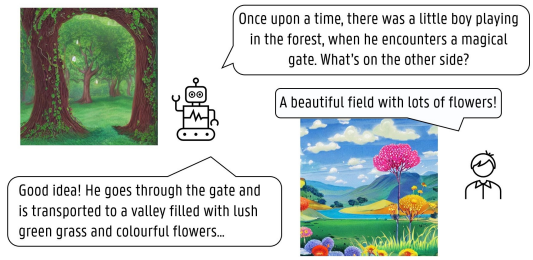

Fig. 2: An example of the interaction between robot and participant. The participant's answer influences the progression of the story.

Additionally, during each turn, the LLM was prompted to add a description of an image fitting to the story. This description was fed to a text-to-image model. This image was shown to the student during the robot's turn, as soon as it is generated. The goal of these images was to increase participants' engagement, ground the interaction in a visual depiction and thereby increase their comprehension of the narrative. To further optimise the users' understanding, on the same screen on which the images are shown, subtitles appeared in synchrony with the robot's speech.

The ASR and text-to-speech are included in the Furhat robot, either using Google Cloud or Microsoft Azure. Here, Google Cloud was used. The images were generated using a Stable Diffusion model running on a local server [15]. GPT-4o was used as LLM, accessed through its API. The resulting

TABLE I: Excerpts of Task Completion Cases

Robot turn before	User answer	Robot turn after	Task completion
... What does he see on the other side of the bridge?	He sees a tiger	Across the bridge, Max sees a beautiful tiger walking near a river. ...	✓
... What clue does Claire find in the forest?	She finds a very large cross	In the forest, Claire finds a hidden cave, ...	✗
... What happens when he tries on the magic shoes?	I don't know	They're not like other shoes! When he tries on the sparkling shoes, ...	N/A
... The map indicates a treasure in the forest, where should she go?	[Unintelligible]	She decides to enter the forest, looking for the treasure ...	Removed for suggestion

architecture of the system that drives the interaction is shown in Figure 3.

C. Data Overview

To evaluate the ASR, audio data was collected. The Furhat robot has a built-in *listen* function, that handles turn-taking. The timing of this function was used to save audio clips, recorded using the Furhat's external microphone. The conversation consisted of 7 robot turns and 6 user turns in between, resulting in 102 audio clips of user speech ($17 \times 6 = 102$). Additionally, as turn-taking errors are not uncommon, especially for inexperienced users, audio was also recorded of the full interaction using the Furhat's microphone. Logs were also kept of the full interaction, including the images, robot speech, user speech transcripts and the messages sent to the LLM. After the interaction, participants were asked to fill out a questionnaire on their experience during the study and to provide demographic data.

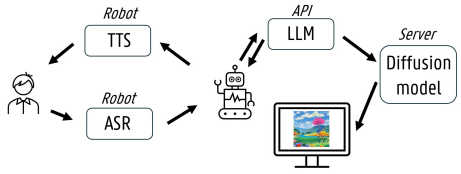

Fig. 3: The architecture of the interaction system, including robot, user interface and a local server.

The Furhat robot's built-in turn-taking mechanism and automatic speech recognition (ASR) occasionally encountered challenges. Specifically, when users delay their response or speak too softly, the system often fails to detect any speech. This issue was observed during the study, where it occurred in more than half of the participants' turns. Therefore, the full audio recordings were cut into clips containing user speech and transcriptions were made using OpenAI's open source ASR system, Whisper.

Other options, such as the Google Cloud or Microsoft Azure ASR systems were omitted, as previous work shows superior performance of Whisper [6]. As seven different Whisper models, each differing in parameter size, are available, the audio was transcribed with each one. Then, the new transcriptions were inserted into the LLM messages in the same way as during the interaction and new answers were generated. This resulted in 714 transcriptions and LLM answers ($17 \times 6 \times 7 = 714$).

Ground truth transcriptions of the audio clips were constructed by the first author, who has a similar language background to the participants and a sufficient knowledge of the French language. To evaluate the task completion metric, for every robot turn - user turn - robot turn combination ($n = 714$), we annotated whether the final robot turn is a relevant and contingent answer to the user's utterance. The examples in Table I show that this can be annotated objectively. Initial data exploration showed that the robot's questions were sometimes overly suggestive, so unusable speech recognition sometimes still led to successful task completion. These cases were identified and excluded for further evaluation ($n = 19$). An example can be found in Table I. Additionally, user speech was in some cases considered to not be useful for task definition ("I don't know"/no speech). These cases were also excluded from the analysis ($n = 161$). As there was some overlap between aforementioned cases, this resulted in a final dataset of $n = 553$. Table I shows examples of successful and unsuccessful task completion, as well as a case of user speech that did not define a task and an overly suggestive question.

III. Results

In this section, the results will be discussed to obtain answers to the research questions. First, we will look at the ASR quality of non-fluent French in comparison to what is reported in literature. Secondly, we will focus on the impact of using an LLM in series with the ASR system on the task completion metric.

The first research question focuses on the ASR quality of L2 French speech compared to the impressive results shown in benchmarks [1]. As mentioned above, near-human performance has been reported as a WER of around 5% [1]. Figure 4 shows the average WER per Whisper model size. The lowest WER, achieved by the Whisper *large* model, is 32.8%, considerably higher than the performance on the aforementioned benchmarks. It is not unreasonable that this could be ascribed to the lower resources of the French language, the imperfect pronunciation and non-fluency of L2 students and the accents of the speakers, which are all combined here. Literature seems to support this conclusion. In [12], Dutch speakers, which is a considerably lower resource language than French (for comparison, there are 9752 hours of French, 2077

hours of Dutch and 438,218 hours of English in Whisper's training data [11]), obtain at most a WER of 8.00% for read speech and 30.70% for dialogue (which is closer to our setting). As expected, L2 speakers have lower performance of 24.80% WER for read speech and 33.80% for dialogue. This is comparable to our non-fluent, L2 speech ASR performance. On the other hand, L2 English speech leads to varying but overall better recognition. In [10], L1 and L2 English speakers' match error rates (MER, which is similar to WER but cannot exceed 100%) from below 5% to around 15% are reported. This cannot be compared to WER directly, but it follows the hypothesis that, even for L2 speakers, English speech is recognised more easily.

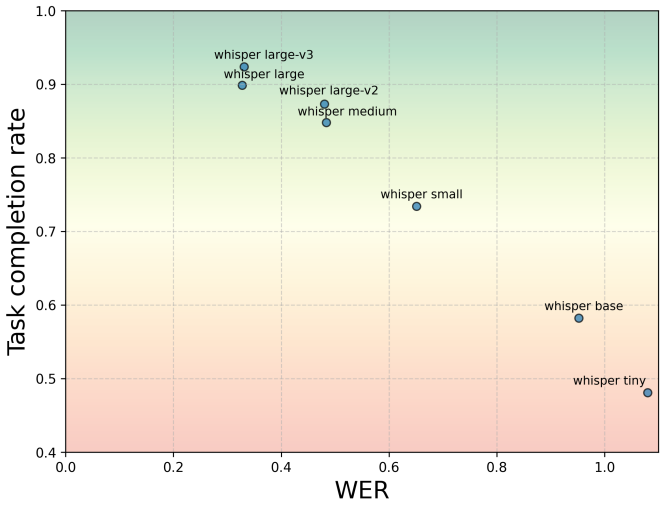

Fig. 4: Scatter plot of Word Error Rate (WER) versus Task Completion rate per Whisper model size. The green color indicates a more successful conversation, while red indicates more communication failures.

The second research question poses whether ASR performance directly influences task completion or, following our hypothesis, if the LLM's pragmatic capabilities allow for task completion even when a non-negligible WER remains. The results discussed above have shown that, although ASR performance has increased considerably in recent years, for non-fluent, L2 French speech the WER is still high, with 32.8% for our dataset using the largest Whisper model, and much higher for small models. Without an LLM to process the resulting transcription, it can be expected that this will lead to task failure most of the time. In Figure 4, task completion is plotted against WER. Here, we found that even with WERs of 32-49%, task completion rates are still between 84 and 92%. Additionally, a Chi-Square test showed significant differences in the task completion of all models ($\chi^2(6) = 77.27$, $p < .001$). Moreover, all models were compared using a pairwise Fischer's exact test. This showed that, among others, the task completion of the three large models and the medium model was not significantly different (p-values ranging from $p = 0.20$ for *large-v3* vs *medium* to $p = 0.81$ for *large-v2* vs *medium*).

IV. Discussion and Conclusion

We studied how well speech recognition works for our specific target group, which combines (i) a lower resource language compared to English, (ii) L2 learner accents and (iii) the non-fluency, imperfect pronunciation and hesitancies typical for language learners. We posed as a hypothesis that this would lead to noticeably worse ASR quality compared to the near-human ASR performance reported for fluent speakers. Additionally, we hypothesised that the emerging pragmatic capabilities of LLMs would allow for high task completion even where ASR performance is lacking. To verify these claims, data was collected during a user study with a group of second language learners. The users took part in a conversation led by a social robot, where each conversational turn defined a clear task where understanding the user's speech was crucial to generate a fitting answer. Turns where no clear task was defined or where the robot's previous turn was considered too suggestive for a fair task evaluation, were left out, resulting in a dataset of $n = 553$ conversational turns and the same amount of user utterance transcriptions.

The audio clips were transcribed using OpenAI's Whisper ASR. This led to a best average WER of 32.8% obtained by the Whisper *large* model, which is considerably worse than performances for typical populations reported in literature. Compared to related research of ASR for non-English speakers, non-fluent speakers and speakers with varying accents, the low performance found here might be explained by both the lower resource language and the effect of L2, non-fluent speakers. Our first hypothesis is confirmed by the data.

To test our second hypothesis, the conversational turns were annotated for their task completion. The task completion rate was compared to WER, the traditional measure of ASR performance. These results, as shown in Figure 4, illustrate that even with a WER as high as 48%, we can achieve a task completion rate of 84%. This confirms our hypothesis that using an LLM in series with an ASR system will cover for enough of the transcription mistakes that we can consider a conversation successful even with high error rates in the transcriptions. Additionally, we found that the four largest Whisper model sizes did not show statistically significantly different task completion rates. Therefore, we can advise the use of the smaller, faster and less energy-consuming Whisper *medium* model, as it shows similar performance in conversation when used in series with an LLM. An additional conclusion that might be drawn from this data, is that current metrics used to evaluate ASR performance do not correctly evaluate this when an LLM is used downstream. As an LLM can use the context of the conversation, common sense and pragmatics, the exact transcription might be less important than recognising the key words in the user's answer, where WER assigns the same importance to every word.

In conclusion, our work demonstrates that while non-English languages and non-fluent speakers pose challenges for ASR, LLM-powered conversational systems are resilient enough to allow for successful interactions.

REFERENCES

[1] W. Xiong, L. Wu, F. Alleva, J. Droppo, X. Huang, and A. Stolcke, "The microsoft 2017 conversational speech recognition system," in *2018 IEEE international conference on acoustics, speech and signal processing (ICASSP)*. IEEE, 2018, pp. 5934–5938.

[2] N. Randall, "A survey of robot-assisted language learning (rall)," *ACM Transactions on Human-Robot Interaction (THRI)*, vol. 9, no. 1, pp. 1–36, 2019.

[3] H. Lee and J. H. Lee, "The effects of robot-assisted language learning: A meta-analysis," *Educational Research Review*, vol. 35, p. 100425, 2022.

[4] Q. Deng, C. Fu, M. Ban, and T. Iio, "A systematic review on robot-assisted language learning for adults," *Frontiers in Psychology*, vol. 15, p. 1471370, 2024.

[5] E. Verhelst, R. Janssens, T. Demeester, and T. Belpaeme, "Adaptive second language tutoring using generative ai and a social robot," in *Companion of the 2024 ACM/IEEE International Conference on Human-Robot Interaction*, 2024, pp. 1080–1084.

[6] R. Janssens, E. Verhelst, G. A. Abbo, Q. Ren, M. J. P. Bernal, and T. Belpaeme, "Child speech recognition in human-robot interaction: Problem solved?" *Companion of the 16th International Conference on Social Robotics - ICSR24*, 2024.

[7] R. Jain, A. Barcovschi, M. Yiwere, P. Corcoran, and H. Cucu, "Adaptation of whisper models to child speech recognition," *arXiv preprint arXiv:2307.13008*, 2023.

[8] A. A. Attia, J. Liu, W. Ai, D. Demszky, and C. Espy-Wilson, "Kid-whisper: Towards bridging the performance gap in automatic speech recognition for children vs. adults," *arXiv preprint arXiv:2309.07927*, 2023.

[9] R. Jain, A. Barcovschi, M. Y. Yiwere, D. Bigioi, P. Corcoran, and H. Cucu, "A wav2vec2-based experimental study on self-supervised learning methods to improve child speech recognition," *IEEE Access*, vol. 11, pp. 46 938–46 948, 2023.

[10] C. Graham and N. Roll, "Evaluating OpenAI's Whisper ASR: Performance analysis across diverse accents and speaker traits," *JASA Express Letters*, vol. 4, no. 2, 2024.

[11] A. Radford, J. W. Kim, T. Xu, G. Brockman, C. McLeavey, and I. Sutskever, "Robust speech recognition via large-scale weak supervision," in *International Conference on Machine Learning*. PMLR, 2023, pp. 28 492–28 518.

[12] S. Wills, Y. Bai, C. Tejedor-García, C. Cucchiarini, and H. Strik, "Automatic speech recognition of non-native child speech for language learning applications," *arXiv preprint arXiv:2306.16710*, 2023.

[13] S. L. Sravanthi, M. Doshi, T. P. Kalyan, R. Murthy, P. Bhattacharyya, and R. Dabre, "Pub: A pragmatics understanding benchmark for assessing llms' pragmatics capabilities," *arXiv preprint arXiv:2401.07078*, 2024.

[14] S. R. Young, "Use of dialogue, pragmatics and sematics to enhance speech recognition," *Speech Communication*, vol. 9, no. 5-6, pp. 551–564, 1990.

[15] R. Rombach, A. Blattmann, D. Lorenz, P. Esser, and B. Ommer, "High-resolution image synthesis with latent diffusion models," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2022, pp. 10 684–10 695.