Studying the Performance of Automatic Speech Recognition Systems on Older Adults.

Carlos Escolano[1], Cristian Barrué[2], Jordi Picas[3] and Guillem Alenyà[2]

***Abstract*— Current Automatic Speech Recognition (ASR) systems still struggle in real-world applications, particularly under challenging noise conditions. In this work, we focus on the case of assistive robots interacting with older adult users. We address this gap by creating a novel evaluation dataset that replicates the acoustic challenges encountered in such scenarios. We benchmark the performance of state-of-the-art ASR systems on this dataset. Our results highlight important limitations when the user uses a monotonous tone, or has speech difficulties, or when the robot is far from the user. Thus, user training in the use of the technology is crucial.**

I. INTRODUCTION

Natural language processing (NLP) plays a crucial role in bridging the gap between human and robot interactions. Ideally, we should be able to converse with robots as effortlessly as we do with humans, without needing any additional interface. To achieve this, robots must be capable of three key processes [1]: understanding the content of a human user's message, generating a logical response based on the message's information, and formatting this response in a grammatically correct manner that is understandable by the user.

In this study, we focus on the first of these steps: evaluating the effectiveness of current Automatic Speech Recognition (ASR) systems in transcribing speech within the context of an assistive robot in a home environment. While modern ASR systems demonstrate impressive performance, they may struggle in a home environment where the robot is mobile. Factors such as background noise and the distance between the speaker and the robot can significantly affect transcription accuracy, potentially rendering current ASR systems unsuitable for such tasks. This issue is particularly pertinent when assisting older individuals, whose specific vocal characteristics may not be well-supported by existing systems. Furthermore, older users may be less familiar with technology, leading to hesitant or unclear interactions.

Our research begins by describing the creation of an evaluation dataset tailored to this task, encompassing various scenarios that simulate real interactions between older users and the robot. We examine different conditions of background noise, microphone distance, and interaction types, noting distinct differences between general and action-oriented speech, as well as the nature of commands issued by older speakers.

We then evaluate Whisper [2], a state-of-the-art multilingual ASR system, to assess its performance across these scenarios. Additionally, we analyze the correlation of ASR results with relevant speaker characteristics such as age, gender, and language used during recordings. This comprehensive evaluation aims to highlight the strengths and limitations of current ASR technology in supporting effective human-robot communication, particularly in assistive contexts involving older users.

II. RELATED WORK

Voice as an interaction modality has been extensively used [3]. However, despite the interest in using voice to communicate with older adults, their efficiency in understanding older adults in languages other than English has not been properly established. One of the main concerns of older adults is about reliability issues that might impede technology adoption [4]. Most approaches use headsets or other systems to avoid most of the issues in capturing the signal. Learning how to use the technology may indeed help. Recent studies demonstrate how older adults adapt over time, and after some training, to voice assistants [5].

One of the primary challenges in Automatic Speech Recognition (ASR) systems is accurately mapping audio utterances to text transcriptions [6], [7], [8]. Traditional ASR systems [9] address this challenge through a series of discrete steps. Initially, audio data is processed to extract acoustic features. These features are then fed into a model that decodes the appropriate sequence of tokens, leveraging language models previously trained on text data.

The rise of deep learning, particularly the Transformer architecture [10], has significantly advanced the field, with end-to-end systems now representing the state of the art [9]. These models consolidate the entire process into a single framework that can extract contextual representations from audio utterances to generate transcriptions. Various end-to-end approaches have been developed. One approach involves Connectionist Temporal Classification (CTC) [11], [12], which exploits the monotonic alignment between audio and transcriptions. Another prominent approach is the encoder-decoder architecture [13], where the decoder functions as a conditional language model, generating transcriptions based on the context provided by the audio's encoder.

[1] Carlos Escolano is with TALP Research Center, Universitat Politècnica de Catalunya, Barcelona, Spain carlos.escolano@upc.edu

[2]Cristian Barrué and Guillem Alenyà are with Institut de Robòtica i Informàtica Industrial, CSIC-UPC, Barcelona, Spain cbarrue@iri.upc.edu, galenya@iri.upc.edu

[3] Jordi Picas is with Suara Serveis, Barcelona, Spain jordipicasv@suara.coop

This work was supported by the project ROB-IN PLEC2021-007859 funded by MCIN/ AEI /10.13039/501100011033 and by the "European Union NextGenerationEU/PRTR"; and the 23S06141-001 FRAILWATCH project funded by Barcelona Ciencia 2020-2023 Plan.

TABLE I

NUMBER OF UTTERANCES FOR EACH SPEAKER AND EACH SCENARIO.

Speaker	Gender	Language	Age	Free-Speech	Command	Read	Noise	Far
speaker1	male	Spanish	78	21	8	15	15	15
speaker2	male	Catalan/Spanish	83	14	9	15	15	15
speaker3	male	Spanish	84	8	9	15	15	15
speaker4	female	Catalan/Spanish	91	15	9	15	15	15
speaker5	male	Catalan/Spanish	67	10	9	15	15	15
speaker6	male	Spanish	86	7	9	15	15	15
speaker7	female	Catalan/Spanish	87	11	9	15	15	15
speaker8	female	Catalan/Spanish	83	10	9	15	15	15

This work focuses on the encoder-decoder approach, utilizing Whisper [2], a state-of-the-art ASR system. Whisper is trained on a large, diverse dataset encompassing multiple languages and varying audio qualities. This extensive training enables the model to be more robust to audio artifacts while achieving state-of-the-art performance in transcription accuracy.

III. DATASET CREATION

To evaluate the performance of the ASR system, we collected data from 8 volunteers aged over 65. All participants, despite having various health conditions, did not suffer from any known cognitive diseases and voluntarily participated in the study. Table I provides an anonymized list of participants, detailing their age, gender, and the language used during the study. The recordings were conducted at the volunteers' apartments, within a municipally-owned assisted community building managed by Suara Serveis, which offers services and activities for the residents. The volunteers live in Barcelona, where it is common to use both Catalan and Spanish, often interchangeably within the same conversation. The procedure has received the approval of the corresponding Ethical committee.

Except for the "Far" setting, all experiments were recorded with a microphone positioned directly in front of the participant at an approximate distance of 15 centimetres. The original recordings were sampled at 44 kHz with a 24-bit rate. Before evaluation, all recordings were downsampled to 16 kHz, which is the expected sample rate for the models.

For this experiment, we devised a script divided into five different parts to simulate various scenarios an assistive robot might encounter in a real home:

a) Free-Speech: At the beginning of the recording, we conducted a 5-minute interview with the participants, focusing on their occupations before retirement and their current everyday lives. This part aimed to make the participants comfortable with the recording process and the researchers, as well as to capture examples of unguided conversations.

b) Command: In this section, we provided a list of 9 different actions, ranging from setting a reminder to asking for an object or requesting help. Participants were asked how they would request a robot to perform these actions. The aim was to understand their natural approach to interacting with the robot without being influenced by provided examples.

c) Read: Participants were given a list of commands corresponding to the actions from the previous section and asked to read them exactly as provided. These recordings served as a reference for the system's performance in terms of distance and environmental sound.

d) Noise: Using the same list of commands, participants were asked to turn on a TV or radio to test the ASR systems' robustness to background noise. They set the volume to their usual level, leading to slight variations in background noise across participants.

e) Far: To simulate a situation where the participant speaks to the robot from a different room, we moved the microphone to the bathroom, the farthest point in the house from the participants' position. The bathroom's acoustics increased reverberation, introducing more artifacts into the recordings to challenge the ASR methods' robustness. Participants then read the 15 commands with background noise, as in the previous scenario.

After obtaining the recordings, we split the original audio into sentence fragments, removing parts where the interviewers addressed the participants. The audio fragments were then manually transcribed to avoid introducing errors that could lead to misleading results if the tested ASR systems produced similar errors.

IV. METHODOLOGY

a) Transcription and Preprocessing: We leverage Whisper large v3[1], a state-of-the-art multilingual ASR system, to generate transcripts for all audio recordings within our dataset. As is common practice in ASR evaluations, both the reference and system-generated transcripts undergo normalization. This involves removing capitalization and punctuation to ensure a fair comparison that focuses solely on word-level accuracy.

b) Word Error Rate (WER): To quantify the performance of Whisper across different scenarios, we employ WER, a well-established metric in the ASR domain. WER represents the edit distance between the reference and system transcripts, normalized by the total number of words in the reference. It is calculated as follows:

$$WER = \frac{S + D + I}{N} \tag{1}$$

where S represents the number of substitutions, D the number of deletions, and I the number of insertions needed

[1]https://huggingface.co/openai/whisper-large-v3

to match the reference. N is the total number of words in the reference transcriptions. Lower WER values indicate better performance, reflecting fewer errors (substitutions, deletions, insertions) in the system-generated transcripts compared to the reference.

V. RESULTS

An inspection of the dataset reveals notable differences in how various scenarios pose challenges to ASR models. Two main patterns emerge from analyzing the average number of words per utterance, as illustrated in Table III. First, free-speech utterances are, on average, three times longer than command utterances. Since ASR methods generate text character by character, longer utterances are more prone to error accumulation, which can be problematic for designing non-guided dialogue systems. Second, the commands proposed by participants are significantly longer than those in the script, suggesting that older individuals address robots differently. This discrepancy highlights potential performance differences between real-world scenarios and controlled laboratory experiments.

Examining the WER results obtained by Whisper, shown in Table II, we observe distinct behaviors for each scenario. For free-speech utterances, the model performs consistently across all speakers, with WER ranging from 0.09 to 0.31 and an average of 0.185. Interestingly, command utterances, despite being much shorter, exhibit similar metrics, with an average WER of 0.186. This may be due to participants' lack of confidence when addressing the robot, leading to hesitation or stuttering.

In the "Read" scenario, where participants read pre-written commands, the average WER drops to 0.10, indicating that ASR systems are more reliable when addressed clearly. When background noise was introduced, the WER only increased by 0.02 on average, demonstrating the system's robustness to noise. Surprisingly, Whisper performed better in this scenario than in the command scenario, underscoring the importance of clear articulation for optimal ASR performance.

Significant differences emerge in the "Far" scenario, where the microphone was placed in a different room. Depending on background noise and individual voice characteristics, WER varied widely. Some participants achieved error rates around 0.2, indicating usable transcriptions, while others experienced severe issues, with the system failing to recognize any original words or generating completely unrelated outputs. Notably, speakers 6 and 7 had error rates exceeding 1, indicating no correct words in the transcriptions.

We also analyzed correlations between WER results and participant demographics, such as age, gender, and language used during the recordings. Table IV shows the correlation between WER and participants' ages across different scenarios. As expected, no significant correlation is observed in the "Read," "Noise," and "Far" scenarios, as all participants read the same sentences. In the "Free-speech" and "Command" scenarios, we observe contradictory results: while free-speech errors positively correlate with age, command errors show a strong negative correlation. This may suggest that older participants have more difficulty recalling past events, but address the robot more directly in the command scenario.

Gender analysis reveals mixed results across all scenarios, as shown in Table V, indicating that the system is robust to both genders. Similarly, language analysis in Table VI shows comparable results for Catalan and Spanish overall. The only exception is the "Far" scenario, where the two cases of severe hallucination occurred in Spanish. However, these differences might be attributed to other factors, such as background noise or difficulties in speaking loudly due to age or health conditions.

VI. DISCUSSION

Beyond recognition performance, our study uncovered interesting patterns in how participants interacted with the robot. Although the script provided did not assign a name to the assistant robot, four out of eight participants spontaneously requested a name to address it. Without prior communication, they independently chose the name "Roberto," a male Spanish name that conveniently includes the "rob" prefix, possibly indicating a tendency to personify the robot.

Another notable pattern emerged in the formulation of commands. The scripted commands were assertive, stating only the action to be performed. In contrast, the commands proposed by participants often included politeness markers such as "please" and "thank you." This led to longer, more elaborate commands, suggesting that older individuals tend to humanize the robot, addressing it as they would a person.

Additionally, we observed the significant impact of prosody on ASR performance, particularly in the "Far" scenario. There was considerable variability in recognition accuracy between speakers. Manual inspection of the audio utterances revealed that participants who used more varied intonation and expressive speech achieved better recognition performance. Conversely, those who spoke in a more monotonous tone or simply read the commands experienced higher error rates and more frequent hallucinations by the ASR system.

These findings highlight the importance of considering user interaction styles and prosody in the design and evaluation of ASR systems, especially for applications involving older users. Understanding these patterns can inform the development of more intuitive and user-friendly human-robot interaction systems.

VII. CONCLUSIONS

In this work, we examined how various scenarios impact the performance of an ASR system used in an assistive robot. Our results indicate that while these systems demonstrate robustness to background noise, they are significantly affected by the distance between the user and the microphone. In distant scenarios, we observed pronounced hallucinations, rendering the system unreliable. Conversely, the system's

TABLE II

WER RESULTS FOR EACH SPEAKER.

Speaker	Gender	Language	Age	Free-Speech	Command	Read	Noise	Far
speaker1	male	Spanish	78	0.1	0.14	0.05	0.08	0.2
speaker2	male	Catalan/Spanish	83	0.27	0.24	0.1	0.1	0.35
speaker3	male	Spanish	84	0.11	0.21	0.04	0.06	0.95
speaker4	female	Catalan/Spanish	91	0.28	0.1	0.16	0.18	0.24
speaker5	male	Catalan/Spanish	67	0.12	0.34	0.13	0.12	0.61
speaker6	male	Spanish	86	0.09	0.13	0.09	0.21	1.06
speaker7	female	Catalan/Spanish	87	0.31	0.22	0.18	0.19	3.9
speaker8	female	Catalan/Spanish	83	0.2	0.11	0.12	0.06	0.21

TABLE III

AVERAGE NUMBER OF WORDS FOR EACH SCENARIO IN DATASET.

Subset	Words/Utterance
Free-speech	32,9
Command	12,8
Read	7,8
Noise	7,8
Far	7,8

TABLE IV

PEARSON CORRELATION BETWEEN AGE AND WER FOR EACH SCENARIO.

Subset	Correlation
Free-speech	0,51
Command	-0,71
Read	0,21
Noise	0,39
Far	0,25

TABLE V

AVERAGE WER RESULTS BY GENDER.

Gender	Free-speech	Command	Read	Noise	Far
Men	0,14	0,21	0,08	0,11	0,63
Women	0,26	0,14	0,15	0,14	1,45

TABLE VI

AVERAGE WER RESULTS BY LANGUAGE.

Language	Free-speech	Command	Read	Noise	Far
Spanish	0,15	0,18	0,09	0,14	1,53
Cat/Spa	0,22	0,20	0,13	0,12	0,35

performance was consistent across gender and language variations, though age introduced some variability, particularly in different scenarios.

Furthermore, our findings highlight the importance of user involvement and approach. When users are familiar with the system and understand how to interact with it, results remain consistent even in challenging conditions. This underscores the critical role of prosodic information in achieving accurate recognition.

The main conclusion from our study is that a combined effort between technical design and user training is essential. Future ASR systems should be tailored to address the specific challenges posed by different interaction scenarios. Additionally, educating users on optimal usage practices is crucial to maximizing the effectiveness of these systems.

REFERENCES

[1] E. Hosseini-Asl, B. McCann, C. Wu, S. Yavuz, and R. Socher, "A simple language model for task-oriented dialogue," in *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual*, H. Larochelle, M. Ranzato, R. Hadsell, M. Balcan, and H. Lin, Eds., 2020.

[2] 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, ICML 2023, 23-29 July 2023, Honolulu, Hawaii, USA*, ser. Proceedings of Machine Learning Research, A. Krause, E. Brunskill, K. Cho, B. Engelhardt, S. Sabato, and J. Scarlett, Eds., vol. 202. PMLR, 2023, pp. 28 492–28 518.

[3] K. Seaborn, N. P. Miyake, P. Pennefather, and M. Otake-Matsuura, "Voice in human–agent interaction: A survey," *ACM Comput. Surv.*, vol. 54, no. 4, may 2021.

[4] A. Pradhan, A. Lazar, and L. Findlater, "Use of intelligent voice assistants by older adults with low technology use," *ACM Transactions on Computer-Human Interaction (TOCHI)*, vol. 27, no. 4, pp. 1–27, 2020.

[5] S. Kim and A. Choudhury, "Exploring older adults' perception and use of smart speaker-based voice assistants: A longitudinal study," *Computers in Human Behavior*, vol. 124, p. 106914, 2021.

[6] M. Giollo, D. Gunceler, Y. Liu, and D. Willett, "Bootstrap an end-to-end ASR system by multilingual training, transfer learning, text-to-text mapping and synthetic audio," in *22nd Annual Conference of the International Speech Communication Association, Interspeech 2021, Brno, Czechia, August 30 - September 3, 2021*, H. Hermansky, H. Cernocký, L. Burget, L. Lamel, O. Scharenborg, and P. Motlícek, Eds. ISCA, 2021, pp. 2416–2420.

[7] C. Escolano, M. R. Costa-jussà, J. A. R. Fonollosa, and C. Segura, "Enabling zero-shot multilingual spoken language translation with language-specific encoders and decoders," in *IEEE Automatic Speech Recognition and Understanding Workshop, ASRU 2021, Cartagena, Colombia, December 13-17, 2021*. IEEE, 2021, pp. 694–701.

[8] P. Duquenne, H. Gong, B. Sagot, and H. Schwenk, "T-modules: Translation modules for zero-shot cross-modal machine translation," in *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing, EMNLP 2022, Abu Dhabi, United Arab Emirates, December 7-11, 2022*, Y. Goldberg, Z. Kozareva, and Y. Zhang, Eds. Association for Computational Linguistics, 2022, pp. 5794–5806.

[9] R. Prabhavalkar, T. Hori, T. N. Sainath, R. Schlüter, and S. Watanabe, "End-to-end speech recognition: A survey," *IEEE ACM Trans. Audio Speech Lang. Process.*, vol. 32, pp. 325–351, 2024.

[10] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin, "Attention is all you need," in *Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017, December 4-9, 2017, Long Beach, CA, USA*, I. Guyon, U. von Luxburg, S. Bengio, H. M. Wallach, R. Fergus, S. V. N. Vishwanathan, and R. Garnett, Eds., 2017, pp. 5998–6008.

[11] A. Graves, S. Fernández, F. J. Gomez, and J. Schmidhuber, "Connectionist temporal classification: labelling unsegmented sequence data with recurrent neural networks," in *Machine Learning, Proceedings of the Twenty-Third International Conference (ICML 2006), Pittsburgh, Pennsylvania, USA, June 25-29, 2006*, ser. ACM

International Conference Proceeding Series, W. W. Cohen and A. W. Moore, Eds., vol. 148. ACM, 2006, pp. 369–376.

[12] M. Gaido, M. Cettolo, M. Negri, and M. Turchi, "Ctc-based compression for direct speech translation," in *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume, EACL 2021, Online, April 19 - 23, 2021*, P. Merlo, J. Tiedemann, and R. Tsarfaty, Eds. Association for Computational Linguistics, 2021, pp. 690–696.

[13] I. Sutskever, O. Vinyals, and Q. V. Le, "Sequence to sequence learning with neural networks," in *Advances in Neural Information Processing Systems 27: Annual Conference on Neural Information Processing Systems 2014, December 8-13 2014, Montreal, Quebec, Canada*, Z. Ghahramani, M. Welling, C. Cortes, N. D. Lawrence, and K. Q. Weinberger, Eds., 2014, pp. 3104–3112.