Dataset and Evaluation of Automatic Speech Recognition for Multi-lingual Intent Recognition on Social Robots

Antonio Andriella
PAL Robotics
Barcelona, Spain
antonio.andriella@pal-robotics.com

Raquel Ros
PAL Robotics
Barcelona, Spain
raquel.ros@pal-robotics.com

Yoav Ellinson
Bar-Ilan University
Tel-Aviv, Israel
yoav.ellinson@biu.ac.il

Sharon Gannot
Bar-Ilan University
Tel-Aviv, Israel
sharon.gannot@biu.ac.il

Séverin Lemaignan
PAL Robotics
Barcelona, Spain
severin.lemaignan@pal-robotics.com

ABSTRACT

While Automatic Speech Recognition (ASR) systems excel in controlled environments, challenges arise in robot-specific setups due to unique microphone requirements and added noise sources. In this paper, we create a dataset of common robot instructions in 5 European languages, and we systematically evaluate current state-of-art ASR systems (Vosk, OpenWhisper, Google Speech and NVidia Riva). Besides standard metrics, we also look at two critical downstream tasks for human-robot verbal interaction: intent recognition rate and entity extraction, using the open-source Rasa framework. Overall, we found that open-source solutions as Vosk performs competitively with closed-source solutions while running on the edge, on a low compute budget (CPU only).

CCS CONCEPTS

• **Computing methodologies** → **Speech recognition**; • **Computer systems organization** → *Robotics*.

KEYWORDS

Automatic Speech Recognition, Audio Dataset, Human-Robot Interaction, Assistive Robotics

ACM Reference Format:
Antonio Andriella, Raquel Ros, Yoav Ellinson, Sharon Gannot, and Séverin Lemaignan. 2024. Dataset and Evaluation of Automatic Speech Recognition for Multi-lingual Intent Recognition on Social Robots. In *Proceedings of the 2024 ACM/IEEE International Conference on Human-Robot Interaction (HRI '24), March 11–14, 2024, Boulder, CO, USA.* ACM, New York, NY, USA, 5 pages. https://doi.org/10.1145/3610977.3637473

1 INTRODUCTION

In recent years, Automatic Speech Recognition (ASR) technology has seen remarkable advancements [16], making its integration into various domains, including human-robot interactions (HRI), increasingly prevalent [18]. ASR systems play a key role in facilitating natural communication between robots and humans, enhancing user experiences, and improving the practicality of robot platforms.

Existing efforts in ASR development and evaluation have primarily focused on datasets collected in controlled or real-world environments such as call centres or voice commands to smart devices, which have significantly advanced ASR technology [10, 12, 14]. Nevertheless, when these ASR systems are implemented on robotic platforms, specific challenges arise. Robot-built-in microphones have unique requirements, such as affordability, non-invasiveness, and ease of integration. Additionally, the recorded sound differs due to additional noise sources, including the robot's internal workstation and its own motors (i.e. while moving its body parts). To overcome these challenges, the research community must create more realistic datasets that address these complexities [17]. Furthermore, it is crucial to understand the extent to which ASR errors can still be deemed "acceptable" for correct classification and responses from the robot. To shed light on these issues, this paper aims to address two fundamental research questions: RQ1) How effective are current ASR systems when applied to robots, and RQ2) To what extent can the input from these systems be used to accurately infer users' intents? Aiming to address these research questions, in this work, we build a multilingual dataset recording the data from two different robot setups (see Fig. 1). Using it as a benchmark, we assess the performance of 4 ASR systems (RQ1). Finally, given the identified sentences, we evaluate how those can be correctly recognised by an open-source intent recognition software (RQ2).

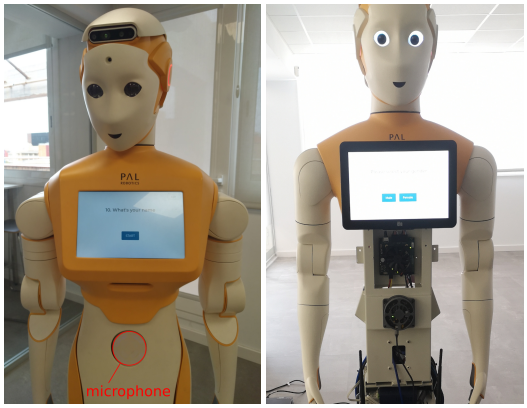
Our findings seek to contribute to the field of HRI, specifically to the development of social and service robots in assistive contexts, fostering a better understanding of ASR capabilities and limitations. This work makes the following contributions: *i*) dataset of sentences in 5 languages with two robot setups; (*ii*) pipeline to evaluate both the ASRs and the impact of a conversational agent (Speech-to-Text and intent recognition); (*iii*) evaluation of the 4 ASRs in 5 languages.

2 EXPERIMENTAL DESIGN

We designed a three-stage experiment. First, we gathered data with the PAL ARI robot [11] to build the audio dataset. Second, we assessed four ASRs on this dataset: Vosk [1], Google Speech [2], OpenWhisper [3] and NVidia Riva [4]. Finally, we evaluated to what extent the text produced by the ASRs is correctly classified

HRI '24, March 11–14, 2024, Boulder, CO, USA

ACM ISBN 979-8-4007-0322-5/24/03…$15.00
https://doi.org/10.1145/3610977.3637473



(a) (b)

Figure 1: ARI robots, (a) with covers and (b) without covers. The microphone is placed underneath the screen.

by the Rasa conversational agent [5] used in the speech pipeline of the PAL Robotics' ARI robot [15].

2.1 Participants

17 adult participants were recruited among Pal Robotics employees mainly (except for one who was visiting the Pal offices at the time of the study), plus 3 children and 1 elderly[1]. The only inclusion criterion was to be a fluent speaker of one of the following 5 languages: English, Spanish, Italian, Catalan and French. 12 self-identified as male, while the remaining 9, as female. The gender distribution was intentionally balanced to maximise equal representation across genders. We asked participants (or their official tutors for the children) to sign a consent form to participate in the study, as well as to make their voice recordings publicly available.

2.2 Data Acquisition

Participants were invited to a meeting room with limited access and requested to interact with the ARI robot. They were asked to indicate their gender (F, M, "don't want to disclose"), their native language and the language they would use to record the data. Next, the data collection was initiated. A sentence was displayed on the screen along with a START button (see Fig. 1). The participant would press the button and read out laud the displayed sentence. A STOP button would immediately appear, so the participant could indicate when they had finished reading the sentence. The next sentence was then displayed. The process repeated until all the sentences were recorded. All sentences appear in the same order.

2.3 Recordings Content

We generated 76 sentences for English, 76 for Spanish, 79 for Italian, 74 for French and 78 sentences for Catalan. The sentences have been grouped based on their assigned intent[6] (i.e., the intent of the sentence), with a total of 28 intents. The dataset is not tailored for a specific task but encompasses a general chitchat used in assistive scenarios. Of those, 23 correspond to general chitchat; 4, to greetings, time and weather; and 1 intent fully related to the content of a specific application of ARI. Some of the sentences contain entities, i.e., particular words we want to extract from a sentence [7]. The sentences correspond to variations of the 28 intents. The reason for the differences in the number of sentences in each language is that some languages allow more or less variations of a given intent.

The sentences recorded by the participants were variable in length, from one-word sentences (e.g., "*Hi*"), to 12-word sentences (e.g., "*Can you explain to me in one sentence what you can do?*") to enrich the dataset. Fig. 2a exemplifies the number of word distributions for the English dataset.

2.4 Apparatus

The recordings were acquired using the ReSpeaker MicArray v2.0 [8], a circular microphone array designed to capture audio from multiple directions. We used two ARI robot setups to gather the data: one with covers (see Fig. 1a) and the other one, without covers (see Fig. 1b). This way, we can also evaluate the ASRs performances in two distinctive conditions: one where the microphone is directly exposed to the participants typical of ASR datasets; and another one where it is enclosed in the robot's covers, which introduces additional sources of noise or distortions, more realistic of ASR datasets in robotic contexts. The microphone is located at the waist level, as shown in Fig. 1.

It's important to note that we are not comparing the performance of the "robot without covers" and the "robot with covers", as the participants for each setup were different. Therefore, the available languages for each sub-dataset also greatly varied. As a result, the dataset provided in this work should be considered as two separate datasets in practice.

2.5 Evaluation Metrics

The goal of the study is to assess not only the performances of the ASRs per se but also the resulting performance of the speech pipeline. Although the ASRs may not provide completely accurate text output, it could still be sufficient for the conversational agent to correctly identify the speech intent, overcoming the limitations of the ASR.

To assess the quality of the ASRs, we employed the following well-known metrics for text analysis: Word Error Rate (WER), Match Error Rate (MER), Word Information Lost (WIL), Word Information Preserved (WIP), Character Error Rate (CER) [9]. However, in this paper, we only present the WER results due to space limitations. The remaining metrics are available in the dataset repository if the reader wants to further analyse them. Moreover, we also evaluate the performance of the intent recognition system by computing the recognition rate of intents and entities from Rasa.

3 RESULTS

To evaluate our dataset, we employed 4 of the most well-known automatic speech recognition systems: Google Speech-to-text, NVidia Riva ASR, Vosk and OpenWhisper. In the case of Google Speech, we do not have access to any specific information about their model, but we can assume that a large model is being used. For NVidia Riva, we used the `conformer-ctc` model [13], which can be considered

[1]The aim of this study is not to compare age groups. However, we included at least two distinct ones to ensure initial diversity in the dataset.

Target	Native	# records	Google Speech			OpenWhisper			Vosk			NVidia Riva		
			WER	Intents	Entities	WER	Intents	Entities	WER	Intents	Entities	WER	Intents	Entities
English	English	302	0.05 (0.16)	0.93	0.90	0.03 (0.16)	0.94	1.00	0.07 (0.21)	0.93	0.82	0.05 (0.15)	0.94	0.90
	French	72	0.08 (0.18)	0.86	0.92	0.10 (0.29)	0.87	0.84	0.12 (0.24)	0.88	0.84	0.11 (0.22)	0.88	0.92
	Italian	151	0.07 (0.19)	0.92	0.84	0.01 (0.05)	0.92	0.96	0.08 (0.19)	0.92	0.84	0.06 (0.16)	0.94	0.92
	Spanish	152	0.23 (0.34)	0.82	0.69	0.07 (0.21)	0.90	0.92	0.26 (0.36)	0.82	0.69	0.14 (0.26)	0.89	0.88
Spanish	Spanish	152	0.05 (0.18)	0.90	0.92	0.05 (0.22)	0.91	0.92	0.09 (0.24)	0.90	0.84	0.06 (0.18)	0.90	0.92
	Italian	151	0.06 (0.17)	0.91	0.88	0.09 (0.22)	0.87	0.84	0.10 (0.24)	0.89	0.92	0.07 (0.19)	0.90	0.84
	French	74	0.16 (0.27)	0.88	0.77	0.20 (0.42)	0.86	0.69	0.23 (0.39)	0.79	0.84	0.18 (0.34)	0.83	0.69
	English	76	0.22 (0.38)	0.83	0.69	0.20 (0.36)	0.86	0.84	0.26 (0.36)	0.82	0.69	0.19 (0.32)	0.81	0.84
French	French	74	0.16 (0.36)	0.86	0.84	0.10 (0.32)	0.90	0.84	0.14 (0.29)	0.85	0.92	0.10 (0.30)	0.87	0.92
	English	72	0.08 (0.18)	0.92	0.85	0.10 (0.29)	0.87	0.84	0.12 (0.24)	0.88	0.84	0.12 (0.30)	0.88	0.84
	Spanish	74	0.20 (0.43)	0.72	0.85	0.27 (0.51)	0.79	0.69	0.24 (0.38)	0.82	0.84	0.24 (0.43)	0.83	0.84
Italian	Italian	236	0.05 (0.16)	0.87	0.82	0.12 (0.31)	0.85	0.84	0.09 (0.25)	0.86	0.84	0.04 (0.15)	0.88	0.89
Catalan	Spanish	155	0.39 (0.41)	0.72	0.85	0.36 (0.47)	0.75	0.76						

Table 1: Performances results in the *without-covers* robot setup for the largest models of the four ASRs. For each of them, we report the M and the SD (in brackets) of the WER and the Rasa intents and entity recognition rate (%).

Target	Native	# records	Google Speech			OpenWhisper			Vosk			NVidia Riva		
			WER	Intents	Entities	WER	Intents	Entities	WER	Intents	Entities	WER	Intents	Entities
English	English	76	0.06 (0.20)	0.93	0.92	0.03 (0.13)	0.93	1.00	0.09 (0.25)	0.90	0.92	0.04 (0.13)	0.93	0.92
	Catalan	304	0.22 (0.33)	0.79	0.52	0.05 (0.17)	0.92	0.82	0.23 (0.35)	0.79	0.53	0.12 (0.21)	0.87	0.63
	Italian	150	0.16 (0.30)	0.85	0.76	0.04 (0.14)	0.90	0.92	0.14 (0.26)	0.88	0.84	0.10 (0.19)	0.88	0.80
Spanish	Spanish	156	0.11 (0.24)	0.90	0.83	0.13 (0.33)	0.89	0.89	0.15 (0.30)	0.89	0.82	0.12 (0.26)	0.90	0.89
	Catalan	303	0.05 (0.17)	0.91	0.87	0.08 (0.22)	0.89	0.86	0.16 (0.30)	0.85	0.80	0.08 (0.23)	0.90	0.86
French	French	74	0.10 (0.28)	0.92	0.92	0.22 (0.47)	0.85	0.84	0.15 (0.35)	0.91	0.84	0.09 (0.29)	0.93	0.92
Italian	Italian	155	0.07 (0.22)	0.86	0.79	0.21 (1.17)	0.85	0.79	0.12 (0.28)	0.85	0.79	0.07 (0.21)	0.87	0.79
	Spanish	79	0.11 (0.22)	0.85	0.77	0.18 (0.29)	0.86	0.69	0.16 (0.31)	0.81	0.69	0.14 (0.21)	0.75	0.46
Catalan	Catalan	388	0.29 (0.40)	0.78	0.78	0.40 (0.54)	0.68	0.78						

Table 2: Performances results in the *with-covers* robot setup for the largest models of the four ASRs. For each of them, we report the M and the SD (in brackets) of the WER and the Rasa intents and entity recognition rate (%).

Target	Native	# records	OpenWhisper			Vosk		
			WER	Intents	Entities	WER	Intents	Entities
English	English	302	0.05 (0.17)	0.92	0.90	0.12 (0.25)	0.90	0.84
	French	72	0.10 (0.20)	0.82	0.92	0.16 (0.23)	0.84	0.92
	Italian	151	0.09 (0.22)	0.87	0.84	0.16 (0.32)	0.91	0.84
	Spanish	152	0.20 (0.32)	0.80	0.69	0.32 (0.36)	0.72	0.73
Spanish	Spanish	152	0.32 (0.43)	0.77	0.80	0.23 (0.33)	0.85	0.73
	Italian	151	0.35 (0.46)	0.82	0.64	0.20 (0.29)	0.87	0.84
	French	74	0.52 (0.46)	0.61	0.38	0.38 (0.43)	0.72	0.46
	English	76	0.62 (0.52)	0.58	0.54	0.46 (0.38)	0.70	0.46
French	French	74	0.63 (0.60)	0.59	0.53	0.26 (0.40)	0.81	0.84
	English	72	0.36 (0.40)	0.74	0.77	0.16 (0.28)	0.85	0.77
	Spanish	74	0.92 (2.01)	0.62	0.46	0.38 (0.42)	0.73	0.62
Italian	Italian	236	0.61 (0.84)	0.68	0.53	0.18 (0.34)	0.84	0.74
Catalan	Spanish	155	0.81 (1.29)	0.52	0.54	0.52 (0.44)	0.67	0.65

(a)

Target	Native	# records	OpenWhisper			Vosk		
			WER	Intents	Entities	WER	Intents	Entities
English	English	76	0.07 (0.19)	0.91	0.85	0.11 (0.24)	0.92	0.77
	Catalan	304	0.16 (0.30)	0.83	0.63	0.32 (0.36)	0.71	0.48
	Italian	150	0.15 (0.28)	0.85	0.84	0.23 (0.30)	0.83	0.80
Spanish	Spanish	156	0.45 (0.84)	0.76	0.69	0.26 (0.32)	0.86	0.79
	Catalan	303	0.43 (0.49)	0.72	0.71	0.33 (0.37)	0.78	0.65
French	French	74	0.68 (1.01)	0.59	0.77	0.17 (0.36)	0.91	0.77
Italian	Italian	155	0.73 (1.82)	0.72	0.46	0.22 (0.38)	0.83	0.67
	Spanish	79	0.73 (1.54)	0.68	0.62	0.29 (0.42)	0.78	0.69
Catalan	Catalan	388	0.79 (0.73)	0.44	0.42	0.58 (0.47)	0.60	0.67

(b)

Table 3: Performances results in the (a) without-covers and (b) with-covers robot setup for the small models of Vosk and OpenWhisper. For each ASR, we report the WER and the Rasa intents and entity recognition rate (%).

a large model. Finally, with respect to OpenWhisper and Vosk, we used respectively the base/small and the large models.

We compared large models for all languages, except for Catalan, since they were not available in either Vosk or NVidia Riva. However, we could only compare small models for OpenWhisper and Vosk. To further evaluate the output of the ASRs and their ability to recognise the sentence's intent, we used pre-trained Rasa models with the dataset's groundtruth sentences in their training. It

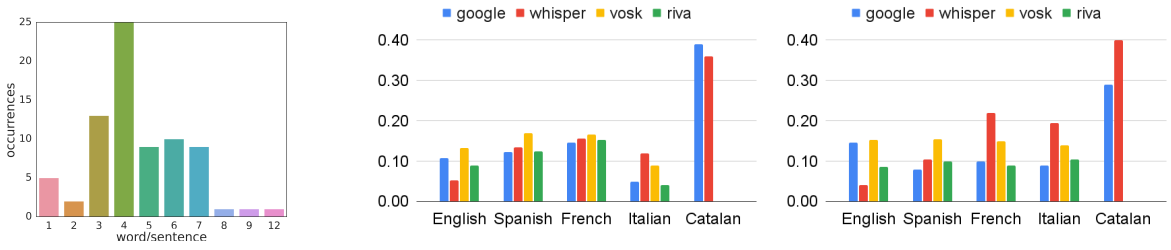


Figure 2: (a) Sentence length distribution for the English dataset. (b) and (c) Average WER performance per target language and ASR in robot without and with covers setups, respectively.

is worthwhile noticing that all the ASRs, except Google ASR, work offline, a key requirement in most of the contexts in which social robots are deployed. The results are reported in Tables 1-2.

ASRs performance for robot without covers. As shown in Fig. 2b, the ASRs achieve overall similar performances within each language, with some noticeable variations in English and Italian. All ASRs poorly perform in Catalan language, most likely evidencing that greater efforts are put into the development of ASRs in commonly spoken languages within Europe. If we look specifically at each language, we can see that OpenWhisper is the best for English and Catalan; Google Speech, for Spanish and French; and NVidia Riva, for Italian. Thus, Google Speech provides good performances across languages, closely followed by OpenWhisper, NVidia Riva, and finally, Vosk performing the worst.

ASRs performance for robot with covers. Similar results were obtained for all the ASRs, except for Catalan (see Fig. 2c). In general, NVidia Riva and Google Speech performed the best, with NVidia Riva being consistently stable across different languages. OpenWhisper achieved good results for English and Spanish, but poor outcomes were obtained for other languages. Regarding the best outcomes for each language, we found that OpenWhisper was the best for English, Google Speech was the best for Spanish, NVidia Riva was closely followed by Google Speech for French, and Google Speech was once again the best for both Italian and Catalan.

Impact of native and not native speakers. We were interested in determining whether there were any differences in speech recognition between native speakers and non-native speakers. To investigate this, we compared the average outcomes of the whole dataset with those of recordings that excluded native speakers. The results showed a slightly lower overall performance, but no significant impact (the results are available in our repository).

Models impact on ASRs performances. In order to determine the impact of model size, we conducted a performance analysis of both Vosk and OpenWhisper's small models. The results are presented in Table 3. We observed that OpenWhisper's performance significantly decreases when moving from the large to the base model, particularly for languages other than English. Furthermore, we noticed that the recognition variance in OpenWhisper was consistently higher than that of Vosk. Overall, Vosk small models outperformed Whisper in all languages except English.

Impact on Rasa intent recognition On average, Rasa can attain correct intent classification ranging from 80% to 90% with all ASRs. However, it seems that even a small error in WER can impact the performance of the Rasa system. For example, when WER was 0.03 (English/English OpenWhisper), we could only get close to 1.

4 DATASET

The dataset associated with this work is stored in a publicly accessible repository: https://osf.io/5kh8g/. Detailed documentation on the usage of the dataset can be found in the same repository. Users are encouraged to download the repository to explore the dataset and access relevant scripts. The dataset will undergo periodic updates to incorporate additions or corrections.

5 CONCLUSIONS

We present a dataset of voice recordings in 5 languages with different robot setups, i.e., robot without and with covers. We evaluated and tested 4 ASR systems: Google Speech, NVidia Riva, Vosk, and OpenWhisper. The purpose of this evaluation is to provide useful insights to future users of ASR systems regarding their performances in relation to different languages, models, and setups.

Regarding RQ1, results indicated that online systems based on large models, such Google Speech achieve superior performances. However, they are also restricted to internet availability, which is not desirable in many applications that require high autonomy of the system. Additionally, their performance comes at the expense of computational resources. In our evaluation, we observed that open-source alternatives like Vosk demonstrate competitive performance compared to closed-source counterparts, particularly when operating on the edge and under a constrained compute budget (e.g., CPU only). Concerning RQ2, as shown in the results, intents recognition can mitigate ASRs error recognition rate achieving acceptable results. Nevertheless, further analysis is needed to understand whether we could make the Rasa model more robust in terms of training examples and learning parameters for small errors of the ASRs.

ACKNOWLEDGMENTS

This project was partially funded by the EU's H2020 research and innovation programme under the Marie Skłodowska-Curie grant agreement No 801342 (Tecniospring INDUSTRY, projects PROCARED and TALBOT) and by the EU H2020 Framework Programme for Research and Innovation SPRING project No. 871245.

REFERENCES

[1] Online. https://alphacephei.com/vosk/.
[2] Online. https://cloud.google.com/speech-to-text/?hl=en.
[3] Online. https://github.com/openai/whisper.
[4] Online. https://www.nvidia.com/en-us/ai-data-science/products/riva/.
[5] Online. https://rasa.com/.
[6] Online. https://rasa.com/docs/rasa/glossary/#intent.
[7] Online. https://rasa.com/docs/rasa/glossary/#entity.
[8] Online. https://wiki.seeedstudio.com/ReSpeaker_Mic_Array_v2.0/.
[9] Online. https://pypi.org/project/jiwer.
[10] Sören Becker, Marcel Ackermann, Sebastian Lapuschkin, Klaus-Robert Müller, and Wojciech Samek. 2018. Interpreting and Explaining Deep Neural Networks for Classification of Audio Signals. *CoRR* abs/1807.03418 (2018). arXiv:1807.03418
[11] Sara Cooper, Alessandro Di Fava, Carlos Vivas, Luca Marchionni, and Francesco Ferro. 2020. ARI: the Social Assistive Robot and Companion. In *2020 29th IEEE International Conference on Robot and Human Interactive Communication (RO-MAN)*. 745–751. https://doi.org/10.1109/RO-MAN47096.2020.9223470
[12] Lingyun Feng, Jianwei Yu, Deng Cai, Songxiang Liu, Haitao Zheng, and Yan Wang. 2022. ASR-GLUE: A New Multi-task Benchmark for ASR-Robust Natural Language Understanding. arXiv:2108.13048 [cs.CL]
[13] Anmol Gulati, James Qin, Chung-Cheng Chiu, Niki Parmar, Yu Zhang, Jiahui Yu, Wei Han, Shibo Wang, Zhengdong Zhang, Yonghui Wu, and Ruoming Pang. 2020. Conformer: Convolution-augmented Transformer for Speech Recognition. arXiv:2005.08100 [eess.AS]
[14] Zohar Jackson, César Souza, Jason Flaks, Yuxin Pan, Hereman Nicolas, and Adhish Thite. 2018. *Jakobovski/free-spoken-digit-dataset: v1.0.8.* https://doi.org/10.5281/zenodo.1342401
[15] S. Lemaignan, S. Cooper, R. Ros, L. Ferrini, A. Andriella, and A. Irisarri. 2023. Open-source Natural Language Processing on the PAL Robotics ARI Social Robot. In *Proceedings of the 2023 ACM/IEEE International Conference on Human-Robot Interaction*. https://doi.org/10.1145/3568294.3580041
[16] Mishaim Malik, Muhammad Malik, Khawar Mehmood, and Imran Makhdoom. 2021. Automatic speech recognition: a survey. *Multimedia Tools and Applications* 80 (03 2021), 1–47. https://doi.org/10.1007/s11042-020-10073-7
[17] Mirko Marras., Pedro A. Marín-Reyes., Javier Lorenzo-Navarro., Modesto Castrillón-Santana., and Gianni Fenu. 2019. AveRobot: An Audio-visual Dataset for People Re-identification and Verification in Human-Robot Interaction. In *Proceedings of the 8th International Conference on Pattern Recognition Applications and Methods - ICPRAM*. INSTICC, SciTePress, 255–265. https://doi.org/10.5220/0007690902550265
[18] José Novoa-Ilic, Rodrigo Mahu, Jorge Wuth, Juan Escudero, Josué Fredes, and Nestor Yoma. 2021. Automatic Speech Recognition for Indoor HRI Scenarios. *ACM Transactions on Human-Robot Interaction* 10 (03 2021), 1–30. https://doi.org/10.1145/3442629