EPFL

CHILI
computer-human interaction
in learning and instruction

LEARN
Center for Learning Sciences

Humans' speech characteristics when interacting with lip-reading social robots

Author:
Estelle Chabanel

Supervisor(s):
Daniel Tozadore

Professor:
Pierre Dillenbourg

January 5, 2023

1 Introduction

As social robots gain importance in our everyday life, the question of our perception and feelings when interacting with them turns out to be an essential subject. Nowadays, lots of researches already studied the topic. Most of them relate people's emotions, trust level, or desire for confidence when talking to a robot. Others analyse people's gestures to reproduce it on robots and make them more "human". However, very few papers were found related human's speech to assert that they are indeed behaving normally, like they usually do. The goal of this study is to investigate this aspect of human-robot interactions.

More particularly, this study relies on a social robot with lip reading ability, implemented as in [6]. An experiment is conducted to explore humans' speech particularities when talking to this kind of robot compared to an usual humans conversation.

2 Literature Review

2.1 Social robots for education

Social robots take an increasingly important role in our society, performing a great variety of functions in every field. In particular, when used for education purposes, social robots can be very useful and make a huge difference. Used as an assistant teacher or an automate friend, researchers and robot developers dream of totally autonomous robots, able to interact independently from any human assistance and discuss with the learners. The possible benefits of such robots have already been proven [5], via numerous experiments like children learning new English words by teaching a care-robot [15] for example. However, the success of such robots not only rely on the robot performance itself but also on the public's reception and their willingness to fully integrate robots in their everyday life [5], [9]. In particular, [9] mentions the importance for children learners of building a socio-affective relationship with their teachers. For robots, this would require certain ability in their social interactions with adapting capability; for humans, it implies a willingness and possibility for faith in their new teaching mate.

Thus, investigating Human-Robot interactions, also referred as HRI is fundamental for the next developments of social robots. HRI studies are now a central subject in research and aim at understanding humans reactions in front of robots to better adapt robots functions and adaptability to their interlocutors.

A brief overview on what could be found about HRI is presented below, to better understand the topic and help design the current experiment.

2.2 Human's-robot interactions

Interactions with social robots contain a lot of different aspects. Taking full advantages of robots potential and minimizing risks requires to study and understand all of them.

Indeed, as illustrated in [7], when talking to robots, people already have initial expectations about their interactions, before they even start talking to the robots. These expectations influence the interactions and have to be taken into account according to the robots utilization. As demonstrated by the experiment, humans have a lower level of liking, anticipate less social presence and have more uncertainty about the upcoming conversation when facing an interaction with a robot than when facing an interaction with a human. Moreover, the aspect of the robot also plays a important role, the more humanly looking it is, the more confident the human feel about his upcoming conversation.

Furthermore, when studying actual human-robot interactions, different measures can be evaluated and reveal various feelings of the human protagonist. Researches focused on gesture [8], dialogues and topics mentioned [11], humans' disclosure [10] and so on. These measures reveal a lot about people's degree of comfort and their willingness to integrate robot in different aspect of their life. Most importantly, it confirmed that difference in treatment of human-human interactions and human-robot interactions can be quantified. However, only few researches were found focusing mainly on the speech and manner of speaking of the humans.

For example, [12] focuses on the adaptation behavior during interaction. Indeed, it has been proven before, that humans constantly try to adapt to their communication partner. As demonstrated by [12], the same occurs when interacting with a robot. These changes are both conscious and unconscious and based both on the prior beliefs of the humans on the robots, and the actual feedback they get during the conversation. The paper quantify this adaptation behavior by subjective entries on the manner of speaking (like clearness of the speech, intonation, loudness...), the structure of the sentences pronounced and the gestures used by the human. However, the experiment only focuses on basic interactions consisting on teaching object names to the robot.

In [8], by looking for indicators revealing the protagonist of the speaker's words, the authors disclosed changes in behavior. Observations show that humans tend to look at robots way more than they look at the other protagonists, even if the addressee is not the robot. It reveals difference in perceptions of the other humans and the robot. Moreover, by showing that acoustic cues can be used to assert whether the participant is talking to the robot or to another human, the paper reveals that some change in speech occurs in HRI, The acoustic cues used are various and not studied in details.

3 Material and Methods

In this context, studying human's behavior is key to predict social robots' success in educational matters. Understanding human's behavior when confronted to robots in its totality will be the main motivation for this study.

Since, studies on people's speech were found missing in the literature review and since efficient lip-reading algorithms, paired with audio recognition model should be the next

big improvement in social robots, the presented experiment focuses on speech characteristics.

Based on the researches presented in section 2, one can wonder if and which differences appear in one's speech when addressed to robot than when addressed to another human. This will constitute the discussed issue of this paper. From our personal experience and the results of existing studies on HRI, the hypothesis are the following:

- One's speech is affected by the addressee, and difference should be observed according to the addressee.
- A person may take a different voice, intonation when addressing to the robot and to other humans.
- A person may consciously or not articulate more and speak slower when addressing to a lip reading robot.

The following section presents the experiment set up to verify or deny these hypothesis. It begins with a brief presentation of the AVSR robot used, before presenting the experiment itself.

3.1 Visual Speech Recognition robot

To study the behavior of humans interacting with robots, a social robot with lip reading ability is used. The lip-reading ability rely on a Visual Speech Recognition (VSR) model developed and presented in [13]. This one was then implemented in a distributed system [6], enabling a Qt robot [3] to process video information of a speaker and translate them into written predictions of the speech using the trained model [13] in a computer GPU.

The use of this distributed system will be precised in the following of this paper.

3.2 Experimental setup

The experiment is divided in two parts. Participants are invited to a brief conversational interaction, both with the robot and an actual "human interlocutor", in order to compare their speech characteristics and be able to assess any differences. The "human-human" interaction will be used as a control group.

For both parts, the conversation script is presented in Figure 1. In both cases, the interaction begins with a presentation of the interlocutor and an invitation for the participant to do the same. By asking for a presentation in both cases, we expect to minimize possible differences caused by the distinct topics since the similarity of the answers in both cases would be the highest as possible.

Then, for the remaining of the discussion, similar topics are chosen for both cases, with only few differences, in order to have comparable answers. The topics are everyday-life related, with questions referring to the participant own habits and tastes. These questions

were chosen in order to be "easy to answer", so the speech characteristics would not be affected by any stress or thoughts latency factors that could be caused by more difficult questions.

Both discussions were video recorded and will be used to evaluate performance of the robot VSR model. However, in the "human-to-human" discussion, the participant was not aware that the video will be used in a VSR model. The audio part of the videos were used to extract some speech characteristics.

	Conversation with the human	Conversation with the robot
Presentation	After a short presentation of the interlocutor What about you ? Can you present yourself ? What are you doing at EPFL?	Hi ! My name is Nao, I am one of the robots of the CHILI lab ! I was born in 2007 and now help the researchers develop new robots. Can you present yourself quickly ? With your name, age, origin ? What are you doing at EPFL ?
Holidays	Can you tell me about your last holidays, what did you do ? did you go somewhere ? with whom ?	What do you like to do on holidays ?
Sports	Do you like watching sports on the TV? Which sports do you usually watch and why do you like it ?	What is your favorite sport ? Why ? How often do you practice it ?
Music	What is your favorite type of music ? When do you listen to it ?	Can you play any instrument ? For how long have you been learning it ? If not, which instrument would you like to play ?
Cooking	How do you make pizza ?	How do you make a cake ?

Figure 1: Script of the experiment

All the participants participated in both conditions in a within-subject experiment where half of them interacted first with the robot and then with the human and the other half the other way around to verify that it did not affect the speech characteristics.

3.3 Measures

Once the experiment recorded, the participants speech behavior were studied and characterized using several metrics.

3.3.1 Speech characteristics

The extraction of the speech characteristics is inspired from [10] and [12] that reported some "manner of speaking" quantitative and subjective measures. The chosen features for the audio analysis were chosen in order to verify the hypothesis. There are listed below.

- The pitch frequency. Humans present wide variation in speech according to their interlocutor, one of the most common and striking one being the intonation. This one can be in part characterized by the frequency of the voice.
- The speech rate. This one is computed thanks to the the number of voice utterance divided by the total time of the participant answers (comprising the pauses).
- The articulation rate. It is computed as the number of voice utterance divided by the total time of speech.

The two last features are extracted using the *my-voice-analysis* Python library that uses duration and sound spectrum [14].

All the presented features are analysed on each answer separately, giving 5 sets of features for each discussion, hence 10 sets for each participants, 2 discussions.

Finally, at the end of the experiments, the participants were asked to fill in a "subjective survey", reporting their feelings about both conversations. This survey is presented in Appendix in Figure 6. The goal of this survey, is to corroborate or add information about the possible change of speech characteristics in the two conversations, via the reporting of the participants' own appreciation on their manner of speaking. However, the survey could also be used to provide first possible explanations for these changes of behavior, the main hypothesis being the comfortability and the unknown situation.

3.3.2 Model performances

The goal of this study also being to check the impact of change of speech on the robot's performance, this last one should be quantified. In this special case, the robot's performance mainly rely on the performance of the VSR model. For each recorded answer, the Words Error Rate (WER) as well as the Character Error Rate (CER) are reported. The computation of both metrics is given in equations (1) and (2) [4],[1].

$$\text{WER} = \frac{\#\text{substitutions} + \#\text{deletions} + \#\text{insertions}}{\#\text{words}} \tag{1}$$

$$\text{CER} = \frac{\#\text{substitutions} + \#\text{deletions} + \#\text{insertions}}{\#\text{characters}} \tag{2}$$

4 First results

10 participants were asked to perform the experiment in order to validate its functioning and observe first results. Hence, if the results presented below can give first intuitions no sure conclusions can be made. We expect the experiment to be performed with more participants in order to be able to verify the first observations.

4.1 Speech characteristics

First, the chosen speech characteristics are extracted for both conversations and each participants, using python tools and the *my-voice-analysis* library as explained in 3.3.1. Codes for analysis of the measures can be found in [2].

To verify the validity of the experiment, the influence of the order of the conversations is studied first. Figure 7 in appendix presents the distribution of the audio features for both types and both orders of the conversations: results for the "human-human" interactions on the left column and "human-robot" interactions on the right one. Comparing both columns, one can observe similar distributions for each features and type of discussions no matter the order. This confirms that the influence of the order of experiments have a negligible impact on the overall results.

The speech characteristics are then studied questions by questions: for each topics, the distribution of the audio features are plotted on Figure 2. The goal of this part is to observe if some important differences can be observed according to the questions topics. As visible, no important variation occur.

The speech rates are distributed in the interval [1, 4], with more occurrences around 3.0; the "human-human" conversation presenting an overall distribution with higher values.

The articulation rates present the bigger variations: they go from 3.0 to 6.0, with distribution shapes quite different according to the questions types. However, the distributions for both types of conversation seem to follow a quite similar shape in all 5 cases.

For the pitch frequency, the obtained values are mostly gathered between 100 and 150 no matter the type of conversations nor the questions topics.

These observations being made, one can look at the distributions for both conversation in a more global way, gathering the different type of questions in order to have more data. The graphs 3, 4 and 5 show the mean features of each participant, gathered on the five subsets of conversations (corresponding to the five questions).

A Table gathering the mean values of the three features for both conversations and all question categories is also presented in 1 for a more quantified observation.

conversation type	mean speech rate	mean articulation rate	mean pitch frequency
human	2.94	4.76	130.94
robot	2.60	4.40	127.68

Table 1: Mean of the speech characteristics for the two types of conversation

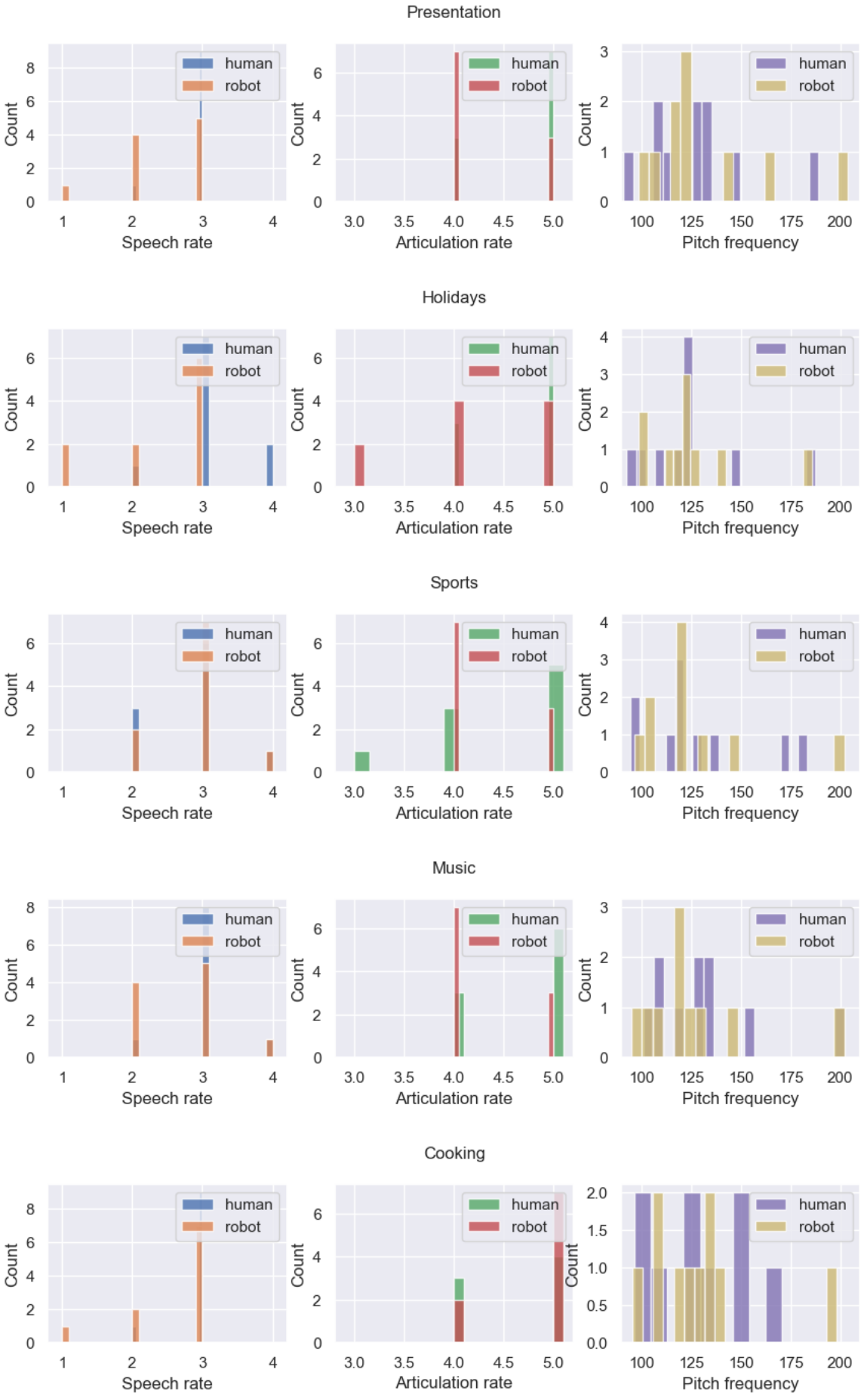

Figure 2: Speech characteristics for each category of questions

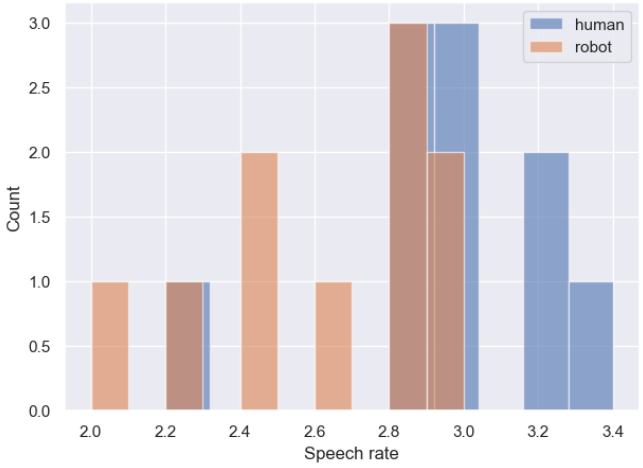

Figure 3: Distribution of the speech rates over the entire conversations

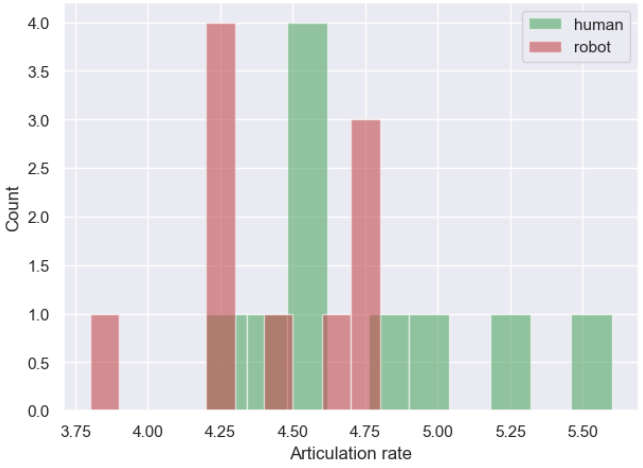

Figure 4: Distribution of the articulation rates over the entire conversations

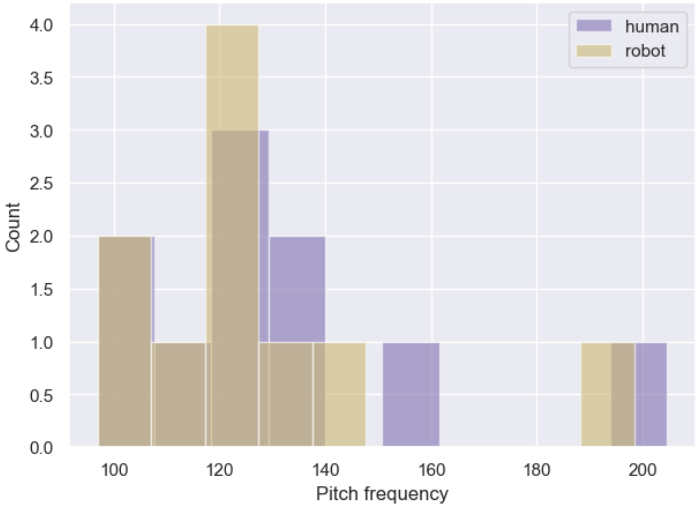

Figure 5: Distribution of the pitch frequencies over the entire conversations

4.2 Model performances

To quantify the performance of the model, recordings of both type of conversations have been evaluated by the VSR model afterwards. However, as the actual model is conceived for very short videos only, lasting 1 to 5 seconds, only recordings of the presentation questions have been used to assess the model's performance. Table 2 below gathers the results of the VSR model on these recordings. The actual predictions are presented in appendix in Figure 8.

participant	conversation type	WER	CER
1	human	0.643	0.556
1	robot	0.4	0.426
2	human	-	-
2	robot	-	-
3	human	0.727	0.5
3	robot	0.889	0.778
4	human	0.737	0.611
4	robot	0.579	0.738
5	human	1.3	0.849
5	robot	0.0	0.0
6	human	0.857	0.745
6	robot	0.714	0.472
7	human	0.615	0.639
7	robot	0.588	0.581
8	human	0.889	0.848
8	robot	-	-
9	human	-	-
9	robot	0.75	0.765
10	human	0.643	0.548
10	robot	0.9	0.708

Table 2: Results of the VSR model

The model failed to evaluate four of the tested videos. By re-analyzing these videos, it was guessed that the framework of the videos, along with the posture of the participants and their gestures did not allow the model to retrieve the mouth landmarks. Indeed, as explained in [6], the model first analyzes the videos' images one by one, find the mouth landmarks and then make predictions that are then balanced using a language model.

The overall performances of the VSR model are not very good, less good than the one presented in [6]. This was expected since the videos are longer, topics covered may be more complex and most importantly, the participants were not aware lip-reading models would be used on the videos, whereas results in [6] were observed on an experiment specifically designed for the VSR model.

However, looking into more details at the model predictions in Figure 8, one can see that for most of the interactions, the model got the overall topic. It even got a perfect prediction for one video (this is not reflected on the WER and CER values due to grammar faults corrected by the model).

5 Discussion, limits and possible improvements

As expected, the speech rate is higher in the "human-human" conversations than in the "human-robot" ones (Figure 3). This was already sensed when performing the experiment and was expected as stated in the hypothesis presented in section 3.

First, since the topics covered in both parts of the experiment are very similar, nearly identical, one can rely on the number of words spoken as a reliable measure of the quantity of information revealed. As visible in the Table 3, participants tend to answer the human with more complete sentences, containing more information than the answer delivered to the robot. With more information to deliver for the same questions, people could unconsciously accelerate their rate of speech not to bore their interlocutor.

	presentation	holidays	sports	music	cooking
human	23.9	41.5	30.8	37.8	51.3
robot	20.9	24.0	33.8	21.3	33.5

Table 3: Mean number of information given for each question category and part of the experiment

Furthermore, this lack of information revealed to the robot could reveal a difference of comfortability with the interlocutor even if the survey did not reveal a huge one. Indeed, when asked if the participants felt comfortable in the conversations, nearly 86% of the participants answer "Very comfortable" for the "human-human" interaction while most of them (71%) qualified the interaction with the robot as "Quite comfortable" only. One could speculate that this gap in the interactions and information revealed is partly due to unconscious feelings and prior expectations of the robot's understandings and behavior. Prior expectations that can also influence the rate of speech, if a participant is aware understanding him takes more processing in the robot than a human head.

Moreover, as revealed by the survey, even if asked similar questions, the participants not only made a difference in the quantity but also in the content of their answers, depending on their interlocutor. Four out of the ten participants reported this observation, participants that began the experiment both with the robot or the human. For example, many of them seemed more confident with the human to make jokes while they were answering more seriously to the robot. This reveals a difference of treatment of the interlocutor, due to prior expectations but also reactions of the robot. However, this difference of treatment is not revealed by the pitch frequencies as speculated in 3. Indeed, the pitch frequency distributions are very similar for both types of interactions (Figure 5), with only slightly higher values for the "human-human" interactions. These slightly higher values are too few to make any speculations, they could be a simple coincidence and interpreting the pitch frequencies would definitely require more data.

Nevertheless, a surprising result is the one obtained for the articulation rate. Unlike our expectations, this one seems to be higher with the human than with the robot.

Possible direction of study to understand this result could be the lack of interest of the participants or the lack of reactions and human-like "posture" of the robot. These observations were noted by several participants in the survey and are some of the limits of the experiment that will be discussed below. One of the participant also mentioned the presence and "non invisibility" of the wizard-of-oz controlling the robot, making it hard to perceive the conversation as a real "human to robot" one and influencing his behavior.

Finally, even if differences in the speech and articulation rates were observed, the performance of the VSR model did not seem to vary much with the types of conversations. This is not surprising considering the experiment set up: many other factors are to take into account in the success of the model, like the gesture, head pose, voice, presence of a beard and so on. The articulation rate, speech rate and pitch frequency are only small factors among many.

Limits and improvements

As mentioned before, several participants noted the lack of natural of the robot conversation. Indeed, the robot was programmed to ask questions but did not react according to the answers making the conversation kind of unnatural. Some participants also mentioned the robotic voice of the robot and the lack of eye contact which make the interaction colder than with the human and could have influence their way of acting and speaking. Almost 90% of the participants acknowledge that they behaved differently with the robot and explained it by the robot's behaviour. Thus, an important improvement to obtain more reliable results would be to program the robot to make it more "human-like", maybe add some reactions and gestures.

Another important thing to note is the background of the participants. As the goal was to only validate the experiment, participants were volunteers from the CHILI laboratory. Thus, most of them had already been in contact and interacted with social robots (7 of them out of 10 exactly). To expand the experiment, reliable results would require participants from different backgrounds, with more of them unfamiliar to social robots as prior expectations and knowledge definitely influence the manner of speaking as explained in [7].

6 Conclusion

This study established a first experiment to explore the differences in manner of speaking of people talking to robot compared to a human-human conversation. It was then linked to performance evaluation of an implemented visual speech recognition robot [6].

First results confirmed some of the hypothesis that is humans present different speech characteristics depending on their interlocutor: they indeed tend to speak slower in front of robots, and may slightly change their tone. The content of the speech also varies consequently, with less information provided to robot interlocutors. Based on participants feed-backs and appreciations, several hypothesis were presented to explain these changes

of behavior: comfortability and lack of human aspect of robots are some of the next topics to investigate further. However, these changes did not impact the VSR robot performance that depend on a lot of different factors. Some other observations were more surprising like the counter-intuitive change of articulation rate, quicker in human-robot conversation. Moreover, several limits and factors were revealed to have possibly influence the experiment and affect the results.

For further investigations and more reliable conclusions, the experiment should include more participants from more various background and take some of the highlighted limits into account. The most important ones are improving the human-robot conversations to make it more natural and adapt the robot behavior with more gesture and oral feed-backs.

References

[1] Character error rate (cer). https://readcoop.eu/glossary/character-error-rate-cer/.

[2] Github repository for the experiment analysis. https://github.com/EstelleChabanel/HRISpeechExperiment.

[3] Qtrobot, humanoid social robot for human ai research & teaching. https://luxai.com/humanoid-social-robot-for-research-and-teaching/.

[4] What is wer? what does word error rate mean? https://www.rev.com/blog/resources/what-is-wer-what-does-word-error-rate-mean.

[5] Tony Belpaeme, James Kennedy, Aditi Ramachandran, Brian Scassellati, and Fumihide Tanaka. Social robots for education: A review. *Science robotics*, 3(21):eaat5954, 2018.

[6] David Roche Daniel Carnieto Tozadore, Pierre Dillenbourg. *Distributed System for Fast and Accurate Visual Speech Recognition*. 2022.

[7] Chad Edwards, Autumn Edwards, Patric R Spence, and David Westerman. Initial interaction expectations with robots: Testing the human-to-human interaction script. *Communication Studies*, 67(2):227–238, 2016.

[8] Michael Katzenmaier, Rainer Stiefelhagen, and Tanja Schultz. Identifying the addressee in human-human-robot interactions based on head pose and speech. In *Proceedings of the 6th international conference on Multimodal interfaces*, pages 144–151, 2004.

[9] Elly A Konijn, Matthijs Smakman, and Rianne van den Berghe. Use of robots in education. *The International encyclopedia of media psychology*, pages 1–8, 2020.

[10] G Laban, JN George, V Morrison, and E Cross. Tell me more! assessing interactions with social robots from speech. paladyn j behav robot. 2021; 12: 136–159.

[11] Min Kyung Lee and Maxim Makatchev. How do people talk with a robot? an analysis of human-robot dialogues in the real world. In *CHI'09 Extended Abstracts on Human Factors in Computing Systems*, pages 3769–3774. 2009.

[12] Manja Lohse, Katharina J Rohlfing, Britta Wrede, and Gerhard Sagerer. "try something else!"—when users change their discursive behavior in human-robot interaction. In *2008 IEEE International Conference on Robotics and Automation*, pages 3481–3486. IEEE, 2008.

[13] Pingchuan Ma, Stavros Petridis, and Maja Pantic. End-to-end audio-visual speech recognition with conformers. In *ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 7613–7617. IEEE, 2021.

[14] Shahab Sabahi. my-voice-analysis. `https://github.com/Shahabks/my-voice-analysis`, 2017.

[15] Fumihide Tanaka and Shizuko Matsuzoe. Children teach a care-receiving robot to promote their learning: Field experiments in a classroom for vocabulary learning. *Journal of Human-Robot Interaction*, 1(1):78–95, 2012.

A Subjective survey

Was it your first interaction with a robot ? *

- Yes
- No

Did you feel comfortable talking to the person ? *

	1	2	3	4	5	
Very uncomfortable	○	○	○	○	○	Very comfortable

Did you feel comfortable talking to the robot ? *

	1	2	3	4	5	
Very uncomfortable	○	○	○	○	○	Very comfortable

Do you think you behave differently with the robot than with the person ? *

- Yes
- No

If yes, in what way ? (speed of speech, articulation, voice intonation, gesture, ...)

Réponse longue

Figure 6: Survey filled by the participants after the experiment

B Speech characteristics results

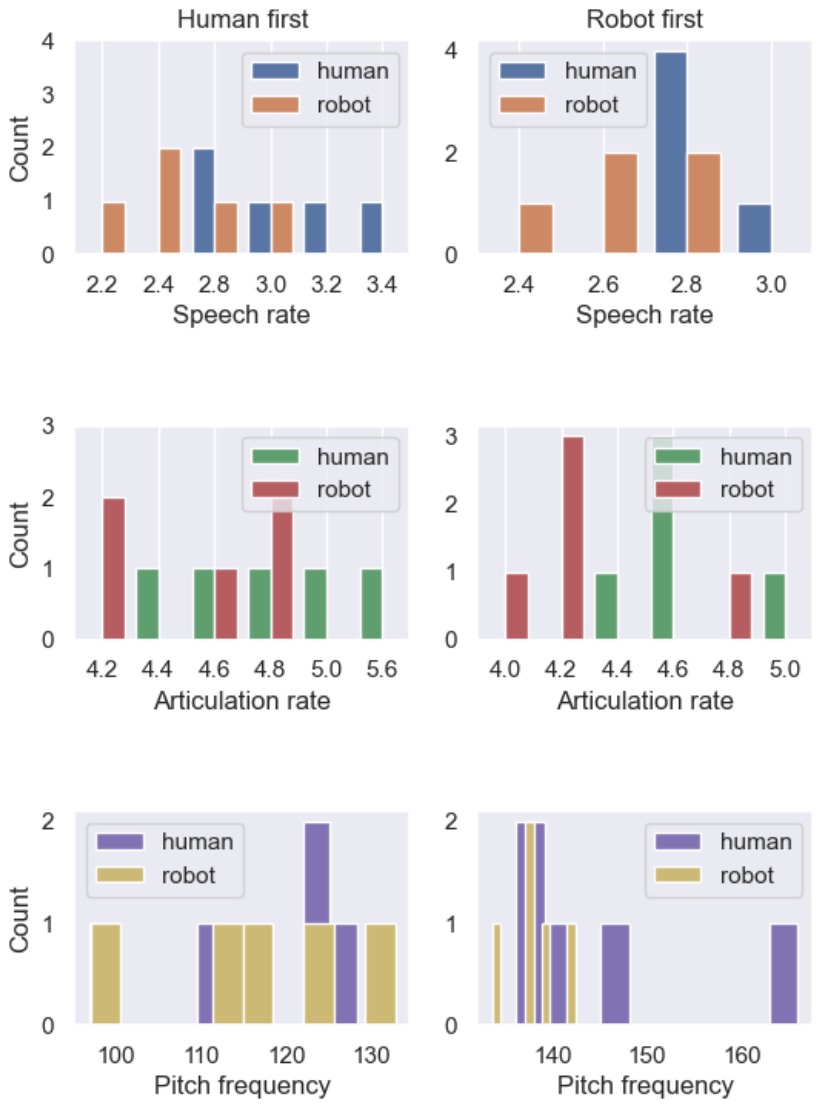

Figure 7: Speech characteristics according to the order of the experiments

C Model performance results

Participant	Conversation type	wer	cer	groundtruth	prediction
1	human	0.643	0.556	DATA SCIENCE AND I AM WORKING HERE AS PART TIME IN THE DINAMILIS PROJECT	AND I AM REALLY HERE BUT I AM IN SOME OF THIS PROJECT
1	robot	0.4	0.426	MY NAME IS KAMIL I AM TWENTY I AM STARTING MY MASTER IN DATA SCIENCE	STILL MY NAME IS CAMEL I AM TWENTY I AM STARTING BY MY STUDENTS IN THE NETHERLANDS
2	human	1	1	YES I AM LAURENT FOURTY FOUR I AM A DEVELOPPER HERE FOR THE DINAMILYS PROJECT	*take and rebuild the video but does not predict anything*
2	robot	1	1	YES SO I AM LAURENT FOURTY FOUR AND I AM A DEVELOPPER AT EPFL	*take and rebuild the video but does not predict anything*
3	human	0.727	0.5	MY NAME IS ANTHONY I AM WORKING ON THE DYNAMILIS PROJECT	AND YOU CAN MAKE SURE THAT I AM WORKING ON THIS PROJECT
3	robot	0.889	0.778	I AM FOURTY ONE AND I AM FROM FRANCE	COME TO THE ONE THAT YOU COME FROM YOUR FRIENDS
4	human	0.737	0.611	I AM TWENTY SIX YEARS OLD I AM ALSO FROM FRANCE AND I AM A P H D STUDENT	I AM REALLY EXCITED TO TALK ABOUT SOME FRIENDS AND I AM COMPETITION
4	robot	0.579	0.738	I AM TWENTY SIX YEARS OLD AND I AM A P H D STUDENT AT E P F L	PERSONALLY I AM TWENTY SIX YEARS OLD AND I AM GOING TO MAKE SURE THAT THATS A LITTLE BIT FUNNY
5	human	1.3	0.849	I'M THIRTY AND I JUST FINISHED MY DOCTORAL STUDY HERE	IM DOING SOME THINGS IN THE REGION OF THE EARTH AND THE EARTH
5	robot	0.0	0.0	I AM THIRTY YEARS OLD AND I HAVE NO RELIGION	I AM THIRTY YEARS OLD AND I HAVE NO RELIGION
6	human	0.857	0.745	HELLO I'M A P H D STUDENT FIRST YEAR AT E P F L	HELLO HELLO IM A MISSION STUDENT AND THIS IS MY THING THAT
6	robot	0.714	0.473	HELLO MY NAME IS JIE I'M A P H D STUDENT AGE THIRTY ONE	HELLO MY NAME IS SIR I AM AN AMERICAN STUDENT AND THE ANSWER IS
7	human	0.615	0.639	MY NAME IS CYRIL I AM TWENTY ONE I AM ORIGINALLY FROM LEBANON	SCHOOL MY NAME IS DAN I AM JUST ONE OF ONE YEARS OLD I AM ORIGINALLY FROM
7	robot	0.588	0.581	HI MY NAME IS CYRIL I AM 21 YEARS OLD I'M A MASTER STUDENT IN DATA SCIENCE	HI MY NAME IS LISTENING I AM TWENTY ONE YEARS OLD AND I AM A MAJOR STUDENT
8	human	0.889	0.848	I AM A MASTER STUDENT AS WELL IN E P F L I AM TWENTY ONE YEARS OLD	I TOLD YOU THAT YOU WERE GOING TO KEEP EVERYTHING BUT WHAT I WANTED TO DO
8	robot	1	1	I AM HAMZA I AM TWENT ONE YEARS OLD I COME FROM MOROCCO I AM A MASTER STUDENT AT E P F L	*take and rebuild the video but does not predict anything*
9	human	1	1	MY NAME IS YOUSSEF I FROM MOROCCO ORIGINALLY I'M TWENTY YEARS OLD	*take and rebuild the video but does not predict anything*
9	robot	0.75	0.765	HI MY NAME IS YOUSSEF I AM 21 22 YEARS OLD ACTUALLY	MY NAME IS SCHOOL OF TWENTY ONE TWENTY TWO HUNDRED AND
10	human	0.643	0.548	MY NAME IS VICTOR I'M FROM MADRID I WORK HERE AT THE CHILI LAB	MY NAME IS MICHAEL I'M FROM MICHIGAN I'M WORKING TO ARTIST I'M A GENERAL
10	robot	0.9	0.708	MY NAME IS VICTOR I'M FROM MADRID I'M THIRTY ONE	HELLO NOW MY NAME IS MICHAEL I AM FROM BETWEEN AND THIS

Figure 8: Predictions of the VSR model on the presentation answers