Gaze and Filled Pause Detection for Smooth Human-Robot Conversations

Miriam Bilac, Marine Chamoux, and Angelica Lim

***Abstract*— Let the human speak! Interactive robots and voice interfaces such as Pepper, Amazon Alexa, and OK Google are becoming more and more popular, allowing for more natural interaction compared to screens or keyboards. One issue with voice interfaces is that they tend to require a "robotic" flow of human speech. Humans must be careful to not produce disfluencies, such as hesitations or extended pauses between words. If they do, the agent may assume that the human has finished their speech turn, and interrupts them mid-thought. Interactive robots often rely on the same limited dialogue technology built for speech interfaces. Yet humanoid robots have the potential to also use their vision systems to determine when the human has finished their speaking turn.**

In this paper, we introduce HOMAGE (Human-rObot Multimodal Audio and Gaze End-of-turn), a multimodal turn-taking system for conversational humanoid robots. We created a dataset of humans spontaneously hesitating when responding to a robot's open-ended questions such as, "What was your favorite moment this year?". Our analyses found that users produced both auditory filled pauses such as *"uhhh"*, as well as gaze away from the robot to keep their speaking turn. We then trained a machine learning system to detect the auditory filled pauses and integrated it along with gaze into the Pepper humanoid robot's real-time dialog system.

Experiments with 28 naive users revealed that adding auditory filled pause detection and gaze tracking significantly reduced robot interruptions. Furthermore, user turns were 2.1 times longer (without repetitions), suggesting that this strategy allows humans to express themselves more, toward less time pressure and better robot listeners.

I. Introduction

Films and science fiction have long imagined robots conversing with us as naturally as humans do. Star Wars' C3P-O, the original social robot for "human-cyborg relations", could deftly converse with people of many nations. More recently, the AI character Samantha from the movie "Her", had rich, meaningful discussions with her human counterpart. Today, these imagined characters are becoming closer to reality. Speech processing has made great strides in the last decade, with interactive robots such as Pepper[1] and voice interfaces such as Amazon Alexa[2], OK Google[3], Siri[4], and Cortana[5] hitting the market.

Yet robots still need improvements to converse as naturally as humans do. One issue, for instance, is that we must speak in a very specific way to interact with robots through speech. We must speak clearly, without hesitation or pauses, preferably without any *"umms"* or *"ah"*s. Unfortunately, according to work by George Mahl, humans emit these kinds of disfluencies on average once every 4.4 seconds [26], pausing to allow themselves time to think, for example. In these cases, a speech system could assume the human has finished speaking, and abruptly interrupt or process an incomplete idea.

Various strategies exist to attempt to address this interruption issue. Many of today's voice services rely on automatic speech recognition and natural language understanding to detect if the user's command is complete[6], thus partially solving the interruption issue. In this case, however, if the user's input is composed of multiple sentences, only the first sentence will be processed; incremental dialog strategies [15] can help here. Other systems simply allow the use of a button to end the speaking turn[7]. In proactive dialogue systems, another way to avoid the issue is to ask specific, non open-ended questions such as "Which colour do you like better, red or blue?". More often than not, it is the human who adapts themselves to the system, speaking in one breath a strung-together sequence that is "perfect", conforming their communication style to the machine's constraints.

A. Turn-Taking in Conversational Analysis

Turn-taking has been studied since the late 1960's as part of human conversation analysis (see, for example, [32]). It includes concepts such as **conversational floor**, which can be "held" or "relinquished" when a speaker continues to speak or ends their speaking turn, respectively.

Another important concept is **overlaps**, when one speaker's speech overlaps with the currently speaking person. Depending on the culture or region, overlaps may occur more or less often when conversing [36]. Some overlaps are *cooperative* [38], for instance as continuation of the interlocutor's speech or backchannels such as "uh huh".

On the other hand, some overlaps are *competitive*, which we call *interruptions* in this paper. Seizing the speaking turn and changing the topic can be associated with displays of power, dominance, and threat [16][41]. As such, it could be important for robots and AI to avoid these overlaps, lest they be perceived as dominating over human speakers.

Filled pauses or **fillers** (used interchangeably in this paper), such as *"uh"* or *"umm"*, are frequent in natural conversation and indicate thinking and/or a desire to continue speaking

*This work was supported by PSPC ROMEO2.

[1]All authors are with SoftBank Robotics Europe, 43 Rue Colonel Pierre Avia, Paris, 75015 France {mbilac, mchamoux, alim}@softbankrobotics.com

[1]http://doc.aldebaran.com/2-5/naoqi/interaction/dialog/aldialog.html

[2]https://developer.amazon.com/alexa

[3]https://developers.google.com/voice-actions/interaction/

[4]https://developer.apple.com/sirikit/

[5]https://developer.microsoft.com/en-us/cortana

[6]https://developer.amazon.com/public/solutions/alexa/alexa-voice-service/reference/speechrecognizer

[7]https://www.google.com/intl/en/chrome/demos/speech.html

[10]. The general consensus in the linguistic community is that these are not errors but a normal part of language and conversation [6], [25], [29], [35]. Relatedly, there exist also silent pauses between words or phrases, separating *installments* of speech within a speaking turn [39].

Gaze is another way to indicate the end of a speaking turn. Studies show that humans will typically look up, to the side, or down while thinking, and then return their gaze to their interlocutor when they are finished speaking [19] [3]. A thorough review of gaze in conversation can be found in [30].

B. Previous Work

1) Filled Pause Detection: In the audio processing field, much study has already been done on detecting filled pauses [40]. In 1999, Goto [18] detected filled pauses in real-time by tracking the fundamental frequency and spectral envelope of speech in Japanese. More recently, thanks to the publication of the Interspeech 2013 SVC dataset, more researchers attempted to detect speech signals such as laughter and filled pauses [23]. For instance, An [2] extracted 9 different features from the speech signal and trained an SVM classifier to detect laughter and filled pauses. Salamin investigated mobile telephone conversations, classifying data into laughter, fillers (including filled pauses), speech or silence [33].

2) Human-Agent Turn-Taking: Some dialogue systems also take into account non-verbal sources of input. For instance, the SimSensei Kiosk by Devault et al. [11] is a state-of-the-art system that allows an animated agent to serve as a virtual therapist. The agent, Ellie, asks open-ended questions such as "Tell me the last time you felt really happy". The authors reported using multimodal information, including gaze detectors, to adapt the agent's non-verbal behaviors. Results based on a subjective questionnaire showed that 56.1% of users would recommend the automated system to a friend. However, no objective results nor details about using gaze detection with respect to turn-taking were reported.

Several human-robot interaction studies have investigated the role of gaze turn-taking. Firstly, some studies adjusted the robot's own gaze to improve human-robot dialogues (see also [24], [5]). Chao studied humans speaking to a robot that spoke in non-linguistic utterances (babbling), and used the robot's gaze to indicate speaking turn [9]. Andrist's study showed that a robot that performed gaze aversion was considered more thoughtful and better managed the conversational floor, with humans interrupting it less [3]. A recent review of robot gaze can be found in [1], and a review about gaze in virtual agents in [31].

In terms of human gaze in human-robot interaction, Sugiyama et al. [37] tracked human speech, face and hip movement to decide if it was appropriate for the robot to take its speech turn. They used Wizard-of-Oz data containing 60 interactions with the NAO robot to perform offline analysis. In that study, they learned that the top salient features when deciding if the robot ought to taking its speaking turn or not were: a) whether the sound it heard was speech or not, and b) user motion and face direction after the sound. The present paper builds on these insights to create, to our knowledge, the first real-time system to use filled pause detection and gaze to manage conversational turn-taking with robots.

C. The HOMAGE System

In this work we present the **HOMAGE** (Human-rObot Multimodal Audio and Gaze End-of-turn) system, built to relax constraints and accept filled pauses, allowing a human to complete their utterance without any time pressure. We create and train a classifier using a new dataset containing filled pauses, captured by a robot's microphones. We also exploit the affordances induced by a humanoid robot face to include the analysis of a human's gaze, captured by a robot's camera, as another cue for speech end-of-turn.

The outline of this paper is as follows. First, we give an overview of the autonomous interactive robot used in this study. Secondly, we describe the creation of a filled pause detection system, as well as its evaluation. We also describe the gaze detector used for this work. Thirdly, we present our HOMAGE turn-taking system that combines our filler detection method and human gaze information into the dialogue system. Finally, we present an experiment evaluating the system. We conclude with discussions and future work.

II. Conversing with Pepper the Humanoid Robot

Our goal in this research is to improve human-robot interaction by detecting and using nonverbal cues of the human to control speaking turn structure. In order to lead a successful conversation, it is important to recognize turn-exchange points and know when it is acceptable or obligatory to take one's turn in conversation, choose an appropriate gap between turns, decide when an overlap is allowed, and so on.

The chosen robotic platform is the humanoid robot Pepper developed by SoftBank Robotics (Fig. 1). The operating system of the robot is NAOqi OS which features a number of modules that allow Pepper to interact with its environment.

In observed human-robot interactions, we noticed that Pepper's dialog module, which endows the robot with conversational skills, contains limited functionality for turn-taking in the case of open-ended questions, such as "How was your summer?". The default solution, where the robot continues the interaction as soon as it detects speech followed by a pause of 200ms, is very efficient for one-word replies like "Good", but in other situations often causes an overlap. For example, Pepper could reply "I'm happy to hear that" before the human finishes giving additional details. Another reason for overlap is that a user may begin their response with a filler such as "hmm..." that may be detected as speech. In these cases, dialogue processing is triggered before the human even started their actual reply. On the other extreme, if the robot waits for several seconds of silence after each utterance before taking its speech turn, humans can be confused due to lack of robot reaction.

III. Filled Pause and Gaze Detection Systems

In this section, we describe the collection of real human-robot dialogue data to 1) use for training of a real-time filled

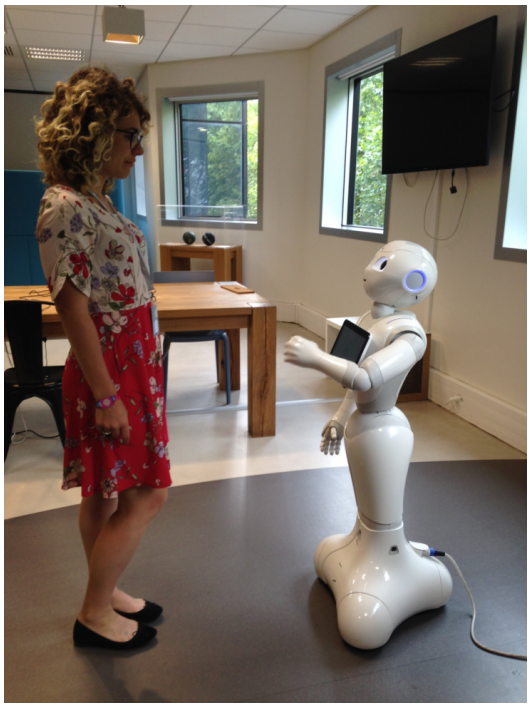


Fig. 1: Conversing with Pepper the humanoid robot.

pause classifier and 2) analyze the pertinence of tracking human gaze for turn-taking. We also provide details about the classifiers and evaluation of the filled pause detector.

Filled pause data collection

We recruited 43 people at SoftBank Robotics Europe with previous experience in robotics. The scenario, developed as a Pepper application using the QiChat dialog scripting language, consisted of open-ended questions about their job satisfaction and ideas for improvements of the robots they worked with. The participants were not informed of the study purpose, asked only to respond to the robot's survey questions.

The participants interacted with Pepper in English. Since some participants did not produce filled pauses, and in some interactions data was not usable because of too much external noise, 31 human-robot interaction formed our final dataset.

Audiovisual data was recorded in the ROSbag format [8] using the NAOqi-ROS bridge[9], useful for collecting multiple data streams from multiple robot sensors in a synchronized fashion. Finally, to train our model for filled pause detection, we annotated 74 filler instances and speech in 168 recorded questions. Only filled pauses of minimum duration of 300 ms, filled with non-verbal utterances like *"uh"* or *"umm"* and appearing throughout different parts of the response, were taken into account. Lexicalized fillers such as "well", "like" or repetitions ("the-the-the") were not included.

[8]http://wiki.ros.org/Bags

[9]http://wiki.ros.org/naoqi_driver

A. Filled Pause Detector

1) Features: Audio analysis was performed on the audio signal recorded from Pepper's microphones. It consists of a four channel interleaved signal with a sample rate of 48000 Hz and 170 ms buffer. There are only slight differences between channels, allowing to perform speech localization. Only front channel signal, from the microphone closest to the speaker, was used for extraction of Mel-Frequency Cepstral Coefficients (MFCC). Cepstral features are often used in speech recognition and showed success in detection of nonlinguistic vocalizations such as laughter [22]. We use 25 MFCC coefficients out of 40 computed Mel bands between 0 and 22050 Hz. Coefficients are computed over a window of 23.2 ms with a 50% frame overlap. To catch the dynamic properties of the signal, i.e., the combination of features over several frames, early temporal integration was performed. The per-frame values for each coefficient were summarized across time using the following summary statistics: minimum, maximum, median, mean, variance, skewness, kurtosis and the mean and variance of the first and second derivatives, resulting in a feature vector of 225 elements. This technique proved to be very efficient in similar problems such as automatic urban sound classification [34].

2) Creation of training and test sets: The dataset was divided into training and test sets, and the former further divided into folds for cross-validation. Folds cannot be generated completely randomly: to avoid artificially high results, slices from same recording should never be placed both in training and test set. Moreover, distribution of filler and speech instances should be constant in each fold.

This distribution problem was solved with a heuristic approach, first-fit decreasing algorithm. First, subjects with the longest filler duration were distributed to each fold and test set. In the next step we tested multiple solutions where we randomly appointed the remaining users in different folds and chose the solution that ensured the most even distribution of fillers.

3) Training: We used the scikit-learn library[10] for Support Vector Machines (SVM) and Random Forest implementations. We experimented with three types of kernel for SVM: linear, polynomial of second degree and RBF. Due to data scarcity we didn't focus on parameter optimization; hyperparameters of the models were set to default values in the library.

All audio features were normalized to a zero mean and unit variance. Both of the models were trained on 77% of data while the remaining 33% were test instances.

4) Classifier evaluation: In this filler-versus-speech classification problem, the dataset was highly imbalanced: approximately 1/4 of the data were fillers and 3/4 were speech. Therefore, a standard evaluation metric accuracy was not appropriate. The chosen evaluation for binary classifiers in this research was area under ROC curve (AUROC), defined on a traditional 0-100% scale and created by plotting the true positive against the false positive rate.

[10]http://scikit-learn.org

Classifier	*AUROC*
Linear SVM	76.2 %
RBF SVM	74.2 %
RF	71.6 %

TABLE I: Evaluation of top performing classifiers. Performance is defined per frame, not filler/speech instances

Cross-validation for optimal window frame for early temporal integration was performed on audio data with 10 lengths in range 0.1 and 1 second. The chosen value was 0.5 seconds. The results of the three best performing models are shown in Table I. Performance is not defined per filler and speech instance but per frame.

This approach is simple but approximates the state of the art: [28] proposes a system for detection of laughter (another type of vocalization) using RASTA-PLP features combined with the temporal derivative and Gaussian Mixture Models for classification, and a similar result of 82.5% is achieved. In [17], a late temporal integration with a weighted average time series smoothing filter using genetic algorithms and an AdaBoost.MH model showed an AUROC score of 89.5%. We believe that a larger training set, as well as further analysis of acoustic properties of filled pauses, with additional information from temporal features, could show even better performance.

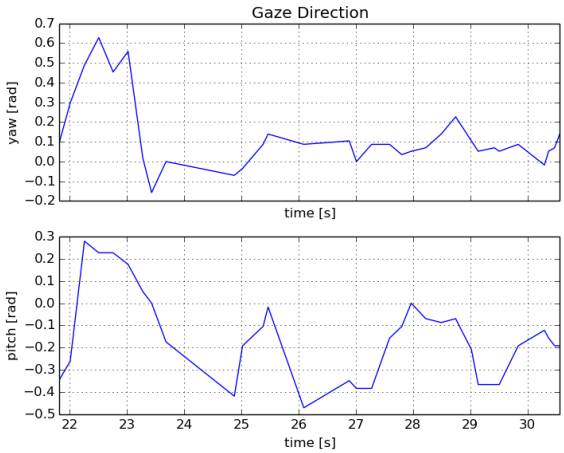
B. Gaze Detector

During the analysis of the present dataset, we noticed that many participants were spontaneously using gaze aversion as another cue to signal turn-holding and turn-yielding. Gaze aversion, a nonverbal cue of cognitive processing, facilitates turn-taking in conversations according to literature in human conversational analysis [19]. People tend to break eye contact at the beginning of the utterance to claim their turn and focus on formulating the answer. At the end of their response, speakers often look at the listener to signal that they finished their answer and that they invite the listener to take the conversational floor [3].

Some HRI studies showed that this human-human gaze aversion behavior is also applicable in human-robot conversation [21], especially for embodied robot such as a humanoid robot [8]. Moreover, we also observed that the same pattern occurred in our recorded human-robot interactions. Therefore, we chose to improve our turn-taking system by adding these cues as input to make the final decision on whether the user is keeping or relinquishing the floor. Pepper's ALGazeAnalysis module was used to collect gaze features with a frequency of 5 Hz: gaze direction, head angles and eye opening degree. In this research we focused on gaze direction.

Figure 2a illustrates the turn-taking behavior of one of the participants while responding to a question. Figure 2b depicts the corresponding gaze direction data, filtered based on the movement of the robot. It is just one example suggesting that humans interact with humanoid robots in a similar way as with other humans. The response starts with a thinking phase accompanied by a rise in gaze direction yaw and pitch

(a)



(b)

Fig. 2: (a) Example of participant's gaze and head movement while responding to a question. Images are taken from the robot's camera during the experiment. (b) Human gaze direction time series plot. Data is recorded from Pepper's ALGazeAnalysis module, where 0 radians corresponds to the human looking into the robot's eyes.

values. The middle phase is the human's verbal reply with occasional pitch value changes. At the end the participant looks back to Pepper, thus giving up the conversational floor in favor of the robot.

Sugiyama et al. [37] highlighted that one of the most important features for turn-taking strategy is the user's face direction after the utterance. Relying on this study, our gaze detector tracks the gaze information only during the most relevant time, i.e. after the end of the utterance. To define the best time frame to collect data, we relied on our own data observation supported by Cathy Pearl's work on end-of-turn timeout [27, p. 113]. According to her, a timeout of 1.5 seconds is a good rule of thumb for voice user interfaces in general. Therefore, we implemented this 1.5s time frame to track gaze direction after the end of a speech utterance, as well as a yaw and pitch value range, to distinguish whether the person is trying to keep or relinquish the conversational floor.

In short, using estimated gaze direction from robot's point

of view collected at the frequency of 5Hz, we defined an angle of +/- 0.15 radians for yaw and pitch, inside which the human is said to be looking at robot, and outside of which the human is said to be averting their gaze. The final decision is made by time averaging data over the given frame.

IV. Multimodal End-of-turn Detection System

In this section, we describe Human-rObot Multimodal Audio and Gaze End-of-turn (HOMAGE) detection system. This is a rule-based model using the Filled Pause Detector and Gaze Detector described in the previous section.

After each user utterance - a continuous speech segment longer than 200 milliseconds - a decision is made whether the user is relinquishing the conversational floor or keeping it.

This decision works as follows:

- Step 1. Filled Pause Detector on heard utterance.
- Step 2. Gaze and audio analysis for 1.5s.
- Step 3. Decision on end-of-turn.

In step 1, audio analysis is made on the utterance with our Filled Pause Detector. The system provides the number of fillers if some occurred in the utterance, along with their position.

In step 2, at the end of the utterance, the system waits for a duration of 1.5s for two purposes. One is to be aware of *installments*, speech separated by silent pauses that can be introduced in spontaneous conversation. Indeed, people can sometimes have short gaps of silence inside their sentence that could be mistaken as the end of turn. To be robust to this situation, we decided to allow any speech detection in this period to restart the system from the step one. The other purpose is to gather gaze information to be processed by our Gaze Detector.

In step 3, after this pause, we make a decision on end-of-turn by combining those cues in a rule-based algorithm. An utterance ending with a filler will automatically lead to a result of *user keeps the turn*. Otherwise, we use the gaze results to make a decision: if the user was looking directly to the robot on average, we presume *user end of turn* and the robot takes the floor. Otherwise, it is a *user keeps the turn* and the robot waits for the user to finish. We implemented these rules in our decision tree (see Figure 3).

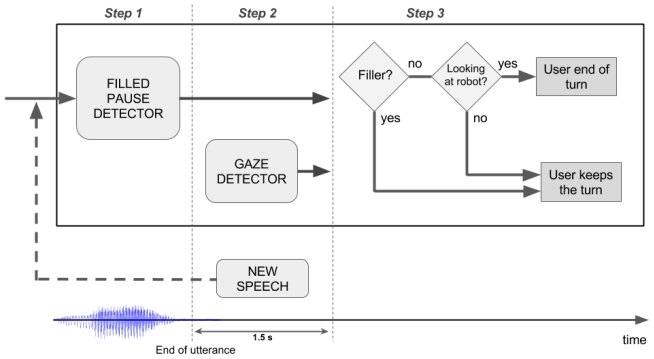


Fig. 3: Overview of the HOMAGE system

V. Experiment and Evaluation

The purpose of this experiment was to test our multi-modal turn-taking system. We implemented the HOMAGE model in the robot Pepper and compared its performance with the default turn-taking model during a dialog interaction.

In this section, we call the default turn-taking model the "gap-turn" system, in which the system considers that the user gives back the floor when the system detects a silent gap longer than 200 milliseconds.

We set up our experiment on two Pepper robots. The idea was to have one robot functioning with the gap-turn system, the other with the HOMAGE system. Both run a questionnaire scenario where the robot poses open-ended questions, as our goal was to produce long answers including hesitations and reflection time. For the participants to not answer the same questions twice, we created two sets of 5 questions each: one set of questions about the company and workplace (Question Set A), and one with more personal questions (Question Set B).

We conducted our experiment on 28 people - 12 females and 16 males - who did not have experience in communication with the robot. Each of them was asked to interact with both robots in English, one after the other. To avoid order and question set bias, we divided them into 4 groups:

- Group 1 (7 participants): Question Set A + Gap-turn system, Question Set B + HOMAGE system
- Group 2 (7 participants): Question Set B + Gap-turn system, Question Set A + HOMAGE system
- Group 3 (7 participants): Question Set A + HOMAGE system, Question Set B + Gap-turn system
- Group 4 (7 participants): Question Set B + HOMAGE system, Question Set A + Gap-turn system

A. Expressiveness of the robot

During conversation, human listeners do not remain still but use backchannels such as nodding or *"uh huh"* sound to express their engagement towards the speaker. Therefore, a contextual backchannel was added to both robots: both robots maintained eye contact and nodded at each end of user utterance. This feedback was aimed to help the user acknowledge that the robot heard them well and to encourage them to speak more.

B. Annotation

Using the software ELAN[11], we annotated each of the participant's responses to each robot question. A human's *response* was defined as the time from the end of the robot's question to the beginning of its next question.

Each response could have two classifications: success and failure.

Success. The response is annotated as a success if the system correctly detected the end of the human's speaking turn. For successful outcomes, we also annotated the duration of the answer of the participant as well as the robot's reaction

[11]https://tla.mpi.nl/tools/tla-tools/elan/

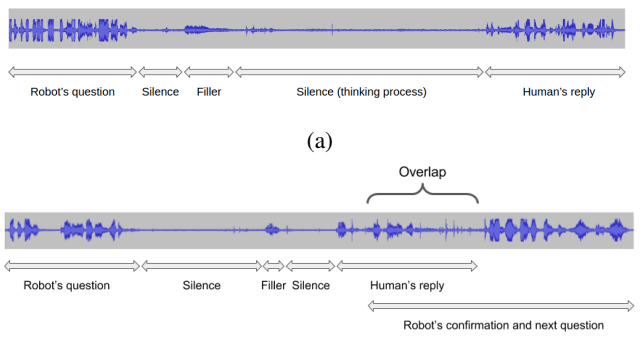


Fig. 4: (a) Successful response. The user ended speaking before robot response. (b) Failed response. The user was interrupted due to a misclassification of a filler followed by a silence as an end-of-turn.

time to the end of their speech. The *answer duration* was defined from the moment the participant started their answer, either with speech or a filled pause, to the end of their verbal answer. The *reaction time* was defined as the time between the last moment the human spoke and the moment the robot started the next question.

Failure. An annotation of failure was given if the participant was either interrupted by the robot (*overlap*) or repeated their answer because of lack of robot reaction (*repetition*). The interruption does not only include situations when the robot and the participant spoke simultaneously, but also when it was evident (based on linguistic meaning) that the participant did not finish their thought. Situations where the user was adding new responses just because the robot did not move to the next question was also considered as repetition.

Examples of successful and failed response in the wave form can be seen in Figure 4.

C. Results

The results among the groups that were presented with the gap-turn model first, and the groups that tested HOMAGE system first were homogeneous. Therefore, we decided to integrate the results of different groups for the same turn-taking strategies. Moreover, due to certain loss of data, only recordings that contained at least 80% of the interaction (answers to minimum 4/5 questions) are included in this quantitative analysis, resulting in 22 human-robot interactions of 4 to 5 questions for the initial turn-taking strategy and 21 interactions of the same number of questions for the HOMAGE system.

Table II depicts the results as a rate of answer outcome to questions, 101 recorded questions for the gap-turn system and 104 questions for the new HOMAGE system. The success rate therefore represents a ratio of questions in which the robot correctly detected the end of the answer (and continued the conversation in appropriate moment), over all recorded questions. The failure rate is split based on the reason of failure: the participant was either interrupted by the robot (overlap) or was forced to repeat the answer. Certain answers can belong to both groups of failure.

		Interaction outcome	
Turn-taking strategy	*Success*	*Overlap*	*Repetition*
Gap-turn System	50.5 %	38.6 %	10.9 %
HOMAGE System	63.5 %	13.5 %	26.0 %

TABLE II: Results of turn-taking strategies as a percentage of the total number of turn exchanges

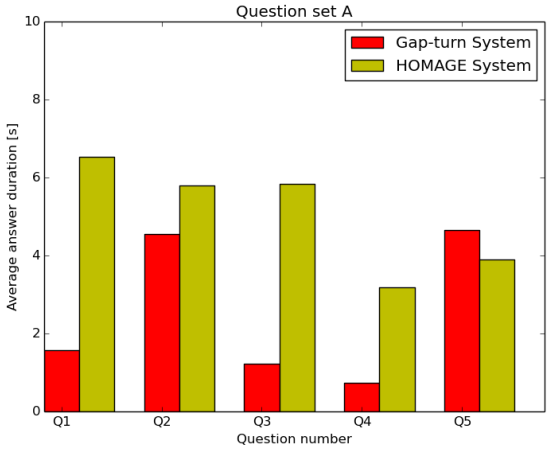
The results show that the new HOMAGE turn-taking strategy decreased the number of interruptions. The difference was found to be statistically significant by Fisher's test at a 95% confidence interval ($p = 5.39 * 10^{-5}$).

Unfortunately, the repetition rate increased ($p = 6.75 * 10^{-3} < 0.05$). One of the reasons can be filler mis-detection that can be addressed by collecting more training data. Moreover, some of the participants did not behave the same way in interaction with the robot as they behave in conversations with other people: some of them felt uncomfortable speaking to the robot and acted disengaged, others averted their gaze from the robot so they could focus more on audio comprehension because of lack of visual feedback (in human-human interaction we gain a lot of information from the lip movement). Thus, those participants didn't provide the necessary gaze feedback to signal that they had finished their turn of speaking. Another reason could be the responsiveness of the systems. Reaction time, defined as the time between the end of the answer of one question and beginning of the next one, was only 0.87 ± 0.26 seconds for the initial gap-turn system and 2.55 ± 0.67 seconds for the new one.

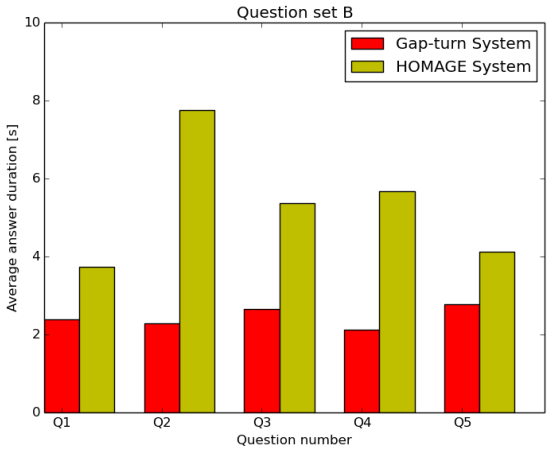
The results also suggest that adding filler and gaze detection increased the success rate. The difference approaches significance but is not statistically significant ($p = 0.0677 > 0.05$). One possible explanation is that tracking gaze direction was not sufficient to detect gaze aversion. Instead, gaze direction should be combined with head angles of the user, as looking significantly away from the robot can result in unknown gaze direction.

Finally, the experiment showed that the new turn-taking strategy increased the average answer duration of users when speaking to the robot (Figure 5). Excluding cases of repetition, when using the HOMAGE system, participants spoke an average of 2.7 seconds longer, an average of 2.1 times longer than the baseline condition.

The results per question are shown in Figure 5, calculated only on successful (non-repetition, non-overlap) responses. The answer to question 5 in Question set A ("How would you like to improve me, Pepper?") is the only one not consistent with the hypothesis. The largest average duration difference of 5.5 seconds can be found in question 2 in the Question set B ("What is your favorite place in Paris and why?") which is a well formulated open-ended question since the user is asked to explain his/her choice.



(a) Question set A



(b) Question set B

Fig. 5: Average answer duration to open-ended questions

VI. Discussion and Future Work

The tendency in interactive technologies, such as social robots, is to use a speech interface to communicate and interact. In this work, we wanted to achieve a step towards natural and fluid spoken conversation between humans and humanoid robots.

In this paper, we argue that gaze is an essential cue in conversational turn-taking. Indeed, our results showed that combining both gaze and filled pause detection led to a decrease in robot interruptions, in the case of the robot asking open-ended questions.

The results also showed that the HOMAGE system resulted in longer human utterances, notably due to this decrease in interruptions. We expect that longer speaker turns, along with strategies such as active listening [20] have the potential to make a robot appear to be more contingent [13] and a better listener [4]. Further user studies should be performed to check the subjective effect of allowing the user to speak longer thanks to HOMAGE.

As the baseline system we chose the default turn-taking model with a silent gap of 200ms, but we believe that a comparison with a system with equal response time can give us even better insight into the importance of gaze and filled pause detection.

In analyzing the recorded sequences, we found out that people also use other cues to keep or release the floor. For example, some of them leaned toward the robot to speak and backed off at the end, as noticed by [37] as well. The intonation of the voice changes too: there could be a rising or falling in voice tone at the end of the last sentence, or a drawn out of the last syllable. Our HOMAGE system could be improved by taking more of those signals into account in a next step.

It is also possible that we could improve the system by understanding the differences between human-human and human-robot relations. When implementing social interaction systems, we often rely on the study of the human-human strategies. Here, we were influenced by sociology research on speech-exchange systems and conversational turn-taking [7] [12], as well as our own understanding of this model. We thus expected the participants to act as in a human-human conversation, but there are obvious differences. For example, sometimes at the end of their speech, people turned their ear toward the robot or closed the eyes to concentrate on the robot answer, while we humans may look at the lip movement of the speaker to better understand their words. Another difference is the robot appearance that may influence the behavior of the participant; Pepper has non-moving eyes plus a tablet. Continued study of human-robot conversation in the wild would be certainly interesting for us to improve our system.

Finally, it would also be interesting to try and understand the impact of such a system in an overall interaction with the robot. Funakoshi et al. [14] provides some insight into the impact of latency; interestingly, it is not necessarily better to provide instant robot replies, but long wait times can make users uncomfortable. Future work should measure the user acceptance of the robot reaction time with the HOMAGE system. The attached video can provide insight into the latency observed during the study.

In particular, we noticed that there appeared to be two types of participants: a) people who speak to the robot as with their phones, adapting their responses to those known turn-taking systems, and b) those who tended to speak in a more natural way, giving longer answers. Reasoning on that, the HOMAGE turn-taking system may have a greater positive impact on the overall feeling of the interaction for those long-speakers. Understanding this bias of the users adaptation to current speech interfaces is an interesting insight for future work.

References

[1] Henny Admoni and Brian Scassellati. 2017. Social Eye Gaze in Human-Robot Interaction: A Review. *Journal of Human Robot Interaction* 7 (2017).

[2] Gouzhen An, David-Guy Brizan, and Andrew Rosenberg. 2013. Detecting laughter and filled pauses using syllable-based features. In *Proceedings of the 14th Annual Conference of the International Speech Communication Association*. 178–181.

[3] Sean Andrist, Xiang Zhi Tan, Michael Gleicher, and Bilge Mutlu. 2014. Conversational gaze aversion for humanlike robots. In *Proceedings of the 2014 ACM/IEEE International Conference on Human-Robot Interaction*. ACM, 25–32.

[4] Gurit E Birnbaum, Moran Mizrahi, Guy Hoffman, Harry T Reis, Eli J Finkel, and Omri Sass. 2016. What robots can teach us about intimacy: The reassuring effects of robot responsiveness to human disclosure. *Computers in Human Behavior* 63 (2016), 416–423.

[5] Dan Bohus and Eric Horvitz. 2014. Managing human-robot engagement with forecasts and... um... hesitations. In *Proceedings of the 16th International Conference on Multimodal Interaction*. ACM, 2–9.

[6] Susan E Brennan and Maurice Williams. 1995. The feeling of another's knowing: Prosody and filled pauses as cues to listeners about the metacognitive states of speakers. *Journal of Memory and Language* 34, 3 (1995), 383.

[7] Justine Cassell. 2000. Nudge nudge wink wink: Elements of face-to-face conversation for embodied conversational agents. In *Embodied conversational agents*. MIT Press, 1–27.

[8] Justine Cassell, Timothy Bickmore, Hannes Vilhjálmsson, and Hao Yan. 2000. More than just a pretty face: affordances of embodiment. In *Proceedings of the 5th international conference on Intelligent user interfaces*. ACM, 52–59.

[9] Crystal Chao and Andrea Thomaz. 2010. Turn Taking for Human-Robot Interaction. In *AAAI Fall Symposium: Dialog with robots*. 132–134.

[10] Nicholas Christenfeld, Stanley Schachter, and Frances Bilous. 1991. Filled pauses and gestures: It's not coincidence. *Journal of Psycholinguistic Research* 20, 1 (1991), 1–10.

[11] David DeVault, Ron Artstein, Grace Benn, Teresa Dey, Ed Fast, Alesia Gainer, Kallirroi Georgila, Jon Gratch, Arno Hartholt, Margaux Lhommet, and others. 2014. SimSensei Kiosk: A virtual human interviewer for healthcare decision support. In *Proceedings of the 2014 International Conference on Autonomous Agents and Multiagent Systems*. International Foundation for Autonomous Agents and Multiagent Systems, 1061–1068.

[12] Starkey Duncan. 1972. Some signals and rules for taking speaking turns in conversations. *Journal of Personality and Social Psychology* 23, 2 (1972), 283.

[13] Kerstin Fischer, Katrin Lohan, Joe Saunders, Chrystopher Nehaniv, Britta Wrede, and Katharina Rohlfing. 2013. The impact of the contingency of robot feedback on HRI. In *2013 International Conference on Collaboration Technologies and Systems*. IEEE, 210–217.

[14] Kotaro Funakoshi, Mikio Nakano, Kazuki Kobayashi, Takanori Komatsu, and Seiji Yamada. 2010. Non-humanlike spoken dialogue: a design perspective. In *Proceedings of the 11th Annual Meeting of the Special Interest Group on Discourse and Dialogue*. Association for Computational Linguistics, 176–184.

[15] Fabrizio Ghigi, Maxine Eskenazi, M Ines Torres, and Sungjin Lee. 2014. Incremental dialog processing in a task-oriented dialog. In *Proceedings of the Fifteenth Annual Conference of the International Speech Communication Association*. 308–312.

[16] Julia A Goldberg. 1990. Interrupting the discourse on interruptions: An analysis in terms of relationally neutral, power-and rapport-oriented acts. *Journal of Pragmatics* 14, 6 (1990), 883–903.

[17] Gábor Gosztolya. 2016. Detecting Laughter and Filler Events by Time Series Smoothing with Genetic Algorithms. In *Proceedings of SPECOM*. Springer, 232–239.

[18] Masataka Goto, Katunobu Itou, and Satoru Hayamizu. 1999. A Real-time Filled Pause Detection System for Spontaneous Speech Recognition. In *Proceedings of Eurospeech 1999*. 227–230.

[19] Simon Ho, Tom Foulsham, and Alan Kingstone. 2015. Speaking and listening with the eyes: gaze signaling during dyadic interactions. *PloS one* 10, 8 (2015), e0136905.

[20] Martin Johansson, Tatsuro Hori, Gabriel Skantze, Anja Höthker, and Joakim Gustafson. 2016. Making Turn-Taking Decisions for an Active Listening Robot for Memory Training. In *International Conference on Social Robotics*. Springer, 940–949.

[21] Martin Johansson, Gabriel Skantze, and Joakim Gustafson. 2014. Comparison of Human-Human and Human-Robot Turn-Taking Behaviour in Multiparty Situated Interaction. In *Proceedings of the 2014 Workshop on Understanding and Modeling Multiparty, Multimodal Interactions*. ACM, 21–26.

[22] Lyndon Kennedy and Daniel Ellis. 2004. Laughter detection in meetings. In *NIST ICASSP 2004 Meeting Recognition Workshop*. 118–121.

[23] Teun F Krikke and Khiet P Truong. 2013. Detection of nonverbal vocalizations using Gaussian Mixture Models: looking for fillers and laughter in conversational speech. In *Interspeech*. 163–167.

[24] Chaoran Liu, Carlos T Ishi, Hiroshi Ishiguro, and Norihiro Hagita. 2012. Generation of nodding, head tilting and eye gazing for human-robot dialogue interaction. In *2012 7th ACM/IEEE International Conference on Human-Robot Interaction (HRI)*. IEEE, 285–292.

[25] Howard Maclay and Charles E Osgood. 1959. Hesitation phenomena in spontaneous English speech. *Word* 15, 1 (1959), 19–44.

[26] George Mahl. 2014. *Explorations in nonverbal and vocal behavior*. Routledge.

[27] Cathy Pearl. 2016. *Designing Voice User Interfaces: Principles of Conversational Experiences*. O'Reilly Media.

[28] Boris Reuderink, Mannes Poel, Khiet Truong, Ronald Poppe, and Maja Pantic. 2008. Decision-level fusion for audio-visual laughter detection. *Machine Learning for Multimodal Interaction* (2008), 137–148.

[29] Ralph Leon Rose. 1998. *The communicative value of filled pauses in spontaneous speech*. Ph.D. Dissertation. University of Birmingham.

[30] Federico Rossano. 2012. Gaze in Conversation. In *The Handbook of Conversation Analysis*, Jack Sidnell and Tanya Stivers (Eds.). John Wiley and Sons, Ltd, Chichester, UK, Chapter 15, 308–329.

[31] Kerstin Ruhland, Christopher E Peters, Sean Andrist, Jeremy B Badler, Norman I Badler, Michael Gleicher, Bilge Mutlu, and Rachel McDonnell. 2015. A review of eye gaze in virtual agents, social robotics and HCI: Behaviour generation, user interaction and perception. In *Computer Graphics Forum*, Vol. 34. 299–326.

[32] Harvey Sacks, Emanuel A Schegloff, and Gail Jefferson. 1974. A simplest systematics for the organization of turn-taking for conversation. *Language* (1974), 696–735.

[33] Hugues Salamin, Anna Polychroniou, and Alessandro Vinciarelli. 2013. Automatic detection of laughter and fillers in spontaneous mobile phone conversations. In *IEEE International Conference on Systems, Man, and Cybernetics*. IEEE, 4282–4287.

[34] Justin Salamon, Christopher Jacoby, and Juan Pablo Bello. 2014. A dataset and taxonomy for urban sound research. In *Proceedings of the 22nd ACM International conference on Multimedia*. ACM, 1041–1044.

[35] Elizabeth Shriberg. 2005. Spontaneous speech: how people really talk and why engineers should care. In *Ninth European Conference on Speech Communication and Technology*. 1781–1784.

[36] Tanya Stivers, Nicholas J Enfield, Penelope Brown, Christina Englert, Makoto Hayashi, Trine Heinemann, Gertie Hoymann, Federico Rossano, Jan Peter De Ruiter, Kyung-Eun Yoon, and others. 2009. Universals and cultural variation in turn-taking in conversation. *Proceedings of the National Academy of Sciences* 106, 26 (2009), 10587–10592.

[37] Takaaki Sugiyama, Kotaro Funakoshi, Mikio Nakano, and Kazunori Komatani. 2015. Estimating response obligation in multi-party human-robot dialogues. In *Proceedings of the IEEE-RAS 15th International Conference on Humanoids*. IEEE, 166–172.

[38] Deborah Tannen. 1983. When Is an Overlap Not an Interruption? One Component of Conversational Style. In *The First Delaware Symposium on Language Studies*, William Frawley Robert J. DiPietro and Alfred Wedel (Eds.). University of Delaware Press, Newark, DE, 119–129.

[39] David R Traum and Peter A Heeman. 1996. Utterance units in spoken dialogue. In *Workshop on Dialogue Processing in Spoken Language Systems*. Springer, 125–140.

[40] Wayne Ward. 1989. Understanding spontaneous speech. In *Proceedings of the workshop on Speech and Natural Language*. Association for Computational Linguistics, 137–141.

[41] Don H Zimmermann and Candace West. 1996. Sex roles, interruptions and silences in conversation. *Amsterdam studies in the theory and history of linguistic science, Series 4* (1996), 211–236.