Gaze Contingency in Turn-Taking for Human Robot Interaction: Advantages and Drawbacks

Oskar Palinko, *Member, IEEE,* Alessandra Sciutti, *Member, IEEE,* Lars Schillingmann, Francesco Rea, Yukie Nagai, *Member, IEEE* and Giulio Sandini, *Member, IEEE*

***Abstract* — It is generally accepted that a robot should exhibit a contingent behavior, adaptable to the needs of each individual user, to achieve a more natural and pleasant interaction. In this paper we have evaluated whether this general rule applies also when the robot plays a leading role and needs to motivate the human partner to keep a certain pace, as during training or teaching. Also among humans, in schools or factories structured interaction is often guided by a predefined rhythm, which facilitates the coordination of the partners involved and is thought to maximize their efficiency. On the other hand, a pre-established timing forces all participants to adjust their natural speed to the external, sometimes not appropriate, timing requirement. Where does the optimal trade-off between these two paradigms lie? We have addressed this question in a dictation scenario where the humanoid robot iCub plays the role of a teacher and dictates brief English or Italian sentences to the participants. In particular we compare a condition in which the dictation is performed at a fixed timing with a condition in which iCub monitors subjects' gaze to adjust its dictation speed. The results are discussed both in terms of participants' subjective evaluation and their objective performance, by highlighting the advantages and drawbacks of the choice of contingent robot behavior.**

I. Introduction

As robots are making their way from factory floors into our everyday lives, the design of their interaction with humans is becoming more and more important. For example in Japan even today it is possible to find robotic greeters when entering electronics stores or mobile service providers (e.g. Pepper). Thus, it is important to pay careful attention to how these new entities will communicate with humans. The basis of interaction is for the robot to respond to the actions of the person: when customers come in, greet them with a smile. As humans naturally use eye contact to establish and modulate interpersonal communication [1] it could be beneficial if robots could also do so while talking to people [2]. Even though a reactive (contingent) approach of the robot is usually favored in conversational turn-taking [3][4] it is still a question if this holds if the role of the robot and its human partners is changed, i.e. what if the robot assumes a leading role, as for example in teaching? Would it be appropriate for it to react to the needs of the "students" or to try to keep a predefined pace not paying attention to reactions? Or more precisely, what amount of contingency is appropriate for a robotic tutor? For example, if we look at a dictation scenario, often present in second-language learning schools, should a potential robot teacher follow the pace of students who are taking notes or should it impose a pace on them? Moreover, to what amount should the robot use eye contact to establish the timing of the interaction?

*Research supported by the European Project CODEFROR (PIRSES-2013-612555)

O. P., A.S., F.R and G.S. Authors are with the RBCS Department of the Fondazione Istituto Italiano di Tecnologia, Genoa, 16163, Italy (corresponding author: +39 01071781475; e-mail: alessandra.sciutti@iit.it).

L.S. and Y.N. Authors are with Graduate School of Engineering, Osaka University 2-1 Yamadaoka, Suita, 565-0871, Osaka, Japan.

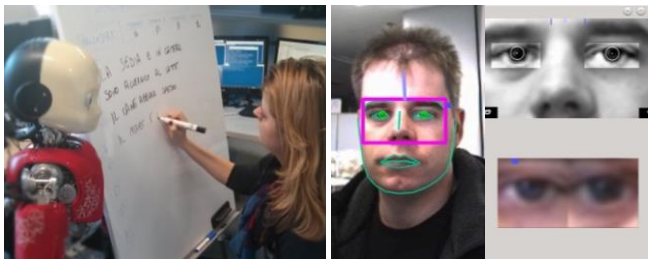

Figure 1. Left: snapshot of the experiment setup; right: output of the mutual gaze detector (pink box – mutual gaze detected).

In our opinion gaze is an important implicit element of communication. For example, humans and more recently even robots can hold their ground in a conversation by just averting their gaze, signaling that they are thinking about what to say [5]. Even more importantly, mutual gaze (i.e. eye contact) detection plays a very important role in turn-taking [6]. In this paper we explore the benefits and drawbacks of augmenting a teaching scenario with implicit gaze communication. We compare a contingent behavior, where the robot reacts to its partner's glances to a purely rhythmic one, where the pace of interaction is preset.

With this task we aim to address two main questions. First, can a mechanism as simple as the detection and the response to subjects' gaze be enough for controlling the turn-taking process in a dictation, with no explicit instructions given to the participants? In other words, can a simple assumption about an automatic interactive behavior - as gazing at the robot to get more information - lead to a working turn-taking system? Second, will the adoption of a responsive or adaptive behavior lead to a more efficient and time-effective interaction avoiding idle times or will it lead to a slower task completion, as participants will tend to slow down when their timing is not regulated by the teacher?

II. Methods

In this experiment subjects assumed the role of students whose task was to write down what their robotic teacher dictated on a whiteboard (see Fig. 1). The robot dictated two sets of 32 short sentences, one in English and one in the participants' mother tongue (Italian). Two different dictating strategies were adopted and presented to the subjects as procedures *alpha* and *beta*. In the *alpha* condition – hereafter *Rhythmic* – the dictation progressed at a predetermine fixed

pace. In the *beta* condition – hereafter *Contingent* – the robot pronounced a sentence only when it detected that the subject was gazing at it, assuming that establishing mutual gaze would signal the readiness of the subject to continue writing. In the following sections, we will provide more details about the system, the different conditions, the subject sample and the data analysis.

A. The Setup

The robot used in the current implementation is the humanoid robot iCub [7]. Our setup leveraged on the use of some existing iCub modules as well as the development of new ones. Fig.2 gives an overview of the system architecture.

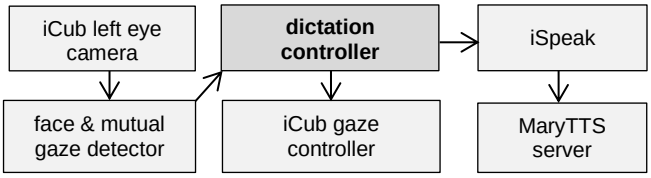

Figure 2. System architecture.

In the following, we will describe each of the elements mentioned above.

1) *iCub left eye camera:* We used the standard iCub camera-grabbing module to acquire videos for our scenario in VGA resolution (640x480) from the iCub's left eye camera (PointGrey Dragonfly2). The speed of image acquisition was around 20 frames per second throughout the experiment.

2) iCub Gaze Controller: This block represents another standard iCub module, iKinGazeCtrl [8], which provides an interface to adjust the robot's gaze direction towards any given point in the camera's image. When a new gaze direction is set, first the eyes perform a saccadic movement towards the goal and then the head turns too, so that the eyes are back to their straight forward direction as much as possible. This capability is employed in our system to track the subject's face with dual intent: a) to keep the face always in the field of view of the camera and b) to provide a more human-like behavior of the robot by directing its gaze towards the subject.

3) *Face and mutual gaze detector:* For detecting the location of the subject's face and the potential mutual gaze in the camera image, we used a face and mutual gaze detection tool, developed by one of the co-authors. The module was initially verified in [9], and now exploited in the current experiment.

The location of human faces is detected by using an open source face detector [10]. Face detection alone is not sufficient to detect mutual gaze. Features corresponding to the location of the pupils need to be extracted from the face region. Furthermore, features correlating to the head pose are required to compensate for different head orientations. First, facial features are extracted using dlib's implementation of [11] which provides robust facial feature detection even on partially occluded faces.

Subsequently the facial features are used to extract the eye regions of each face. On each region, a pupil detection algorithm is applied. The algorithm uses multiple heuristics to generate a candidate map of the most probable pupil location even in the presence of low resolution and noise. First, a gradient-based approach is used to extract center candidates [12]. Second, the region is adaptively thresholded into two luminance classes. The weights for the map corresponding to darker areas are increased. The assumption here is that darker areas more likely correspond to the pupil. Finally, the location of the maximum in the candidate map provides the desired pupil location. A subset of the facial features and the pupil coordinates are used to build a 6-dimensional feature vector. To detect mutual gaze we trained an epsilon-insensitive support vector regression model to estimate the horizontal gaze direction. As training data we used the Columbia gaze database [13]. Using a regression model has the advantage that it is now possible to detect mutual gaze simply by thresholding the estimated horizontal gaze direction. In our experiment, we set a mutual gaze threshold of ±10°.

Once the features are extracted and classified, the mutual gaze status is sent together with the coordinates of each face to the Dictation Controller, which in turn provides the center coordinates of one face to the Gaze Controller module every second, to ensure that the robot keeps a steady gaze on the human participant.

4) *iSpeak:* iSpeak is a standard iCub module which provides speech synthesis functionality to the robot. In our setup it receives textual sentences from the Dictation Controller and passes them on to the MaryTTS module. At the same time it produces simulated lip movements using the LED lights representing the robot's mouth.

5) *MaryTTS:* It is an open-source text-to-speech platform which transform textual sentences into speech using different voices [14].

6) Dictation Controller: The DC module was specifically created for the current experiment. It accepts as input the location of the subject's face and the presence or absence of mutual gaze. As output it sends textual sentences to iSpeak for execution and also tells the iCub gaze controller which way to turn the robot's visual attention. In the *Rhythmic* condition of the experiment the robot does not react to mutual gaze, rather the sentences are sent out to iSpeak with fixed timing. None the less, the robot waits for the issuance of the next sentence proportionally to the length of the previous sentence, which is being written down by the human subject. The waiting time was selected to simulate an average writing time of about 26 words per minute [15]. On the other hand, during the *Contingent* condition, the next sentence is not started until the subject glances back at the robot, after finishing writing. This glance back is the mutual gaze signal sent by the Mutual Gaze Detector module. We require the mutual gaze signal to be present continuously for at least 150ms for it to be recognized as a gaze back event. As an additional constraint we disabled reactions to gazes back at the robot in the first 5 seconds after the end of speech, in order to suppress false positives, as it was impossible to finish writing within such a short period.

B. Subjects

Eight subjects (6 women and 2 men, ranging in age from 26

to 33 yr, mean age 28 yr) took part in the experiment. All subjects were healthy and did not present any neurological, muscular, or cognitive disorder. All participants gave written informed consent before testing. The study was approved by the local ethics committee and all experiments were conducted in accordance with legal requirements and international norms (Declaration of Helsinki, 1964).

C. Procedure

The whole task consisted of the dictation by the humanoid robot iCub of four paragraphs, each composed of 8 short predefined sentences (e.g., "The flowers are red."), in two sessions: one in English and one in Italian. In total subjects had therefore to write 64 short sentences. In particular, for each language, participants encountered two blocks of each condition (*Rhythmic* and *Contingent*) in counterbalanced order (i.e., either R-C, C-R or C-R, R-C). Also the order of language (Italian or English) presentations was counterbalanced among participants to control for order effects. Subjects were instructed to listen to each sentence and then write it down, while leaving blank spaces for any word that they did not understand. The sentences were chosen so that, in each paragraph, the average length was about 19 characters, both for the English and for the Italian sessions. The difference between conditions was that in the *Rhythmic* condition, the robot waited for a fixed time after each sentence (see Sec. IIA), while in the *Contingent* condition, the robot did not initiate a new sentence until the subject gazed at it. In both conditions though the robot moved its head and eyes to look at the subject. The task lasted on average about half an hour per subject and was fully recorded both through the camera in the robot left eye and through an external camera. After the experiment subjects were requested to complete a short questionnaire where they had to rate each of the two procedures (*alpha* or *beta*) on three 7-point scales with respect to the perceived probability to make an error, the pleasantness of the procedure and its difficulty. Then, they were asked to indicate which of the two would have belonged to a more advanced language course and to briefly explain which was the actual difference between them.

D. Data Analysis

The video recordings of all subjects were annotated in ELAN, to individuate the timing of subjects' writing, robot dictating and potential system failures or subjects' strange behaviors (e.g., a posteriori corrections of previously written sentences) [16]. The annotations were then imported in MATLAB through the SALEM Toolbox [17] where they were further analyzed with custom routines. The main variables for the analysis were Task Duration – the time interval between the beginning of the dictation of two subsequent sentences; Wait Time – the time interval between the completion of writing and the beginning of the dictation of the next sentence; the Writing Speed and the Number of Errors in the writing. Furthermore from the responses to the questionnaire we derived a measure of the perceived pleasantness of the two procedures and an evaluation of how clear the understanding of robot behavior in the two conditions was.

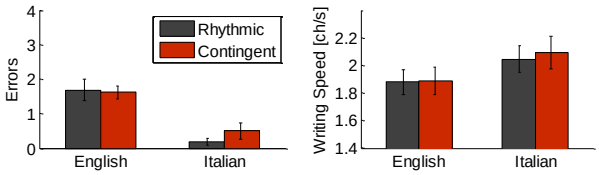

Figure 3. Left: average number of errors per condition; Right: average writing speed per condition. Error bars represent sample standard error (SEM).

III. RESULTS

A. System Errors

During the execution of the experiment there were three cases when a technical error in the Dictation Controller algorithm caused the robot to pronounce two sentences one right after another. Since each time two sentences were affected, we needed to eliminate 6 sentences out of the total 512 (1.17%) from the final analysis. Furthermore, the Gaze Detection algorithm sometimes caused false positives (detected mutual gaze when the subject was not looking at the robot) and false negatives (didn't detect when the subject was in mutual gaze). False positives occurred 7 times, while false negatives were recorded 9 times. Except one time, these false detections did not cause automatic cancellation of the sentences, as they were not disruptive to the process. One time a subject moved out of the field of view of the robot, thus the robot gaze had to be manually redirected back to the subject, which caused one sentence to be eliminated.

B. Subjective evaluations

The first goal of our experiment was to assess whether subjects could perform the dictation in the *Contingent* condition without any explanation of how it worked. All subjects but one automatically adapted appropriately to the task, naturally gazing at the robot after finishing writing. The single exception, who initially stared continuously at the whiteboard, started looking back at the robot after being invited by the experimenter to "interact with iCub", and from that moment on established an appropriate gaze pattern for the rest of the experimental session. To evaluate whether participants had explicitly understood how the system worked in the two different conditions (called generically *alpha* and *beta* during the experiment), in a questionnaire we asked them to describe the difference. Of the 8 subjects only two realized that such difference consisted in how the robot timed its utterances. Of the other 6, three did not describe any difference, while three erroneously thought that a difference existed in the type of sentences used or in the robot voice. It is important to note that although most subjects did not realize that robot behavior in the *beta* (*Contingent*) condition was responsive to their gaze, they naturally exhibited a gaze behavior which was appropriate to guarantee the continuation of the task.

The second question we were interested in was whether a contingent, more adaptable robotic behavior could result in a more pleasant interaction for the human partner. To address this question we asked subjects to evaluate separately the two conditions *alpha* and *beta*, choosing a value on 7-point scales for the probability to make an error, the pleasantness of the task and the difficulty of the condition. Although most subjects could not detect the actual difference between the

two conditions, 5 out of 8 participants found *beta* (*Contingent*) less difficult and less likely to cause errors, and 2 of them found it also more pleasant. The other participant rated the two conditions equally. Accordingly, when asked which of the two conditions would have been part of a more advanced language learning course, the majority (5 over 8) picked *alpha* (2 indicated *beta* and one indicated both).

C. Quantitative Analysis

To be sure that we were not causing additional difficulty to the task with the introduction of the *Contingent* behavior, we compared the number of errors in writing (misspell and blanks) as a function of the condition and the language of the dictation. As it can be seen in Fig. 3 left, the number of errors was significantly higher for English than for Italian ($F(1,7) = 35.48$, $p<0.001$, Two-way Within Measures ANOVA, with Language and Condition as factors), but no difference was present as a function of Condition ($p=0.54$), nor any interaction between Condition and Language ($p= 0.38$). Therefore, the English dictation qualifies as a more difficult task than the Italian one (mother tongue) for our sample, while gaze contingency has no effect on the number of errors. This is confirmed also by an analysis of the average writing speed (Fig. 3, right), which appears to be significantly slower for English than Italian ($F(7,1) = 10.51$, $p = 0.014$), but stable across conditions (Condition: $p = 0.35$; interaction: $p = 0.54$, Two-way Within Measures ANOVA, with Language and Condition as factors).

Since subjects adopted different strategies to cope with difficulty at understanding the dictation in both the *Contingent* and the *Rhythmic* conditions, with some immediately leaving a blank and some spending a long time thinking to the possible completion, we decided to remove the sentences containing errors from the following analyses to reduce inter-individual variability.

An further important question that we wanted to address with our task was whether leaving the possibility to the subjects to – implicitly – control the timing of the dictation could have led to slacking, i.e., to the adoption of a slower pace, especially for those subjects who naturally tend to be slower at writing. To verify this we measured for each subject the time to complete a single sentence (Task Duration), as the time between the beginning of the dictation of one sentence and the beginning of the next. In the *Rhythmic* condition this value was fixed to a predetermined average value (see Methods). In the *Contingent* case, it was determined by when the participant looked back at the robot. In Fig. 4 we plotted individual Task Durations in the *Contingent* condition averaged over each block of 8 sentences (top panel – Italian, bottom panel – English). From the graphs we can derive two observations. First, on average task duration in the *Contingent* condition did not differ significantly from the reference (*Rhythmic*) value. This is confirmed also from a Two-way Within measures ANOVA on the Task Duration averaged between the two blocks, with Language and Condition as factors, where neither factors nor

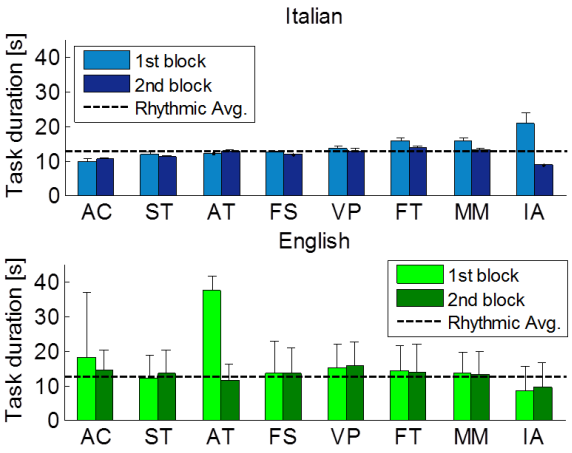

Figure 4: Average Task Duration in the two blocks of the *Contingent* condition for each subject. Error bars represent SEM.

interaction reached significance ($F(1,7)$, $p = 0.46$, $p=0.12$ and $p=0.27$ respectively). Hence, even when implicitly allowed to freely pace themselves, subjects maintained on average the fixed timing predicted by assuming an average writing speed. The second observation regards instead the difference between the first and second blocks of sentences. Indeed, a few subjects (three for Italian and one for English) exhibited a clear adaptation between the two blocks, with a substantial decrease in average Task Duration, up to 140% in Italian and 224% in English. Therefore, at least a subset of subjects needed a few trials to get entrained in the appropriate behavior for the *Contingent* condition to run.
To discount the impact of training on performance, in the following analyses we considered only the second block for all conditions.

Another useful variable to compare the performances and strategies adopted by participants in the two different conditions is represented by the Wait Time, i.e. the time between the completion of the writing of a sentence and the beginning of the dictation of the next. In the *Rhythmic* condition, the dictation timing was fixed, therefore the Wait Time indicates how appropriate the chosen velocity for the subject was, with negative Wait Time implying that the dictation was too fast (the robot started dictating before the subjects completed their writing) and large positive Wait Time indicating that the Rhythm was too slow, potentially leading to boredom and loss of time. In Fig. 5 (left panel) we have plotted individual Wait Times during the last block of each *Rhythmic* condition, as a function of subject's average writing speed. As expected, Wait Time tends to increase with subjects' speed (linear fit slope: 2.10 ± 0.82 (SD), $R^2 = 0.52$ for Italian, slope: 4.51 ± 1.90 (SD), $R^2 = 0.48$ for English). However, for most subjects it is positive and not too long (about 2 seconds), indicating that the timing selected for our *Rhythmic* condition was reasonable for the task at hand.

We then moved to check what happens when the dictation rhythm is not fixed but depends on subjects' gazing. Will slower subjects take more time to process and check what

they have written? Or will faster subjects compensate the short time spent writing by a lengthier check of their sentences? Fig. 5 (right panel) seems to suggest the opposite, i.e. a tendency to converge on average to the same Wait Time (again around two seconds) independently on the average subject's writing speed (linear fit slope: -0.27 ± 0.70 (SD), R^2 = 0.02 for Italian, slope: 0.29 ± 1.82 (SD), R^2 = 0.004 for English). This implies for instance that the faster participants exploited the contingent scenario to accelerate the rhythm of the dictation. The slowest participant, instead, who was often interrupted in his writing in the *Rhythmic* condition, in the *Contingent* case exhibited a slightly slower rhythm that guaranteed him at least a brief time between one sentence and the next.

As a last analysis we evaluated whether task difficulty had an impact on how the same subject dealt with the possibility to control the turn-taking in the interaction. To this aim we compared the Wait Time each subject adopted in the *Contingent* condition with respect to the Wait Time he or

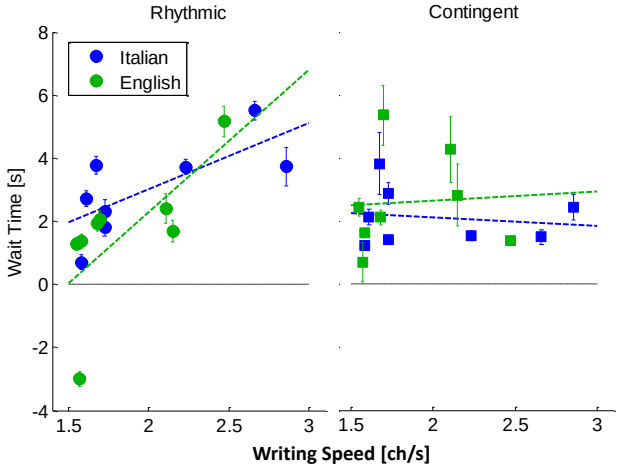

Figure 5: Individual Wait Times as a function of Writing speed. Error bars represent SEM over the last block of each

she exhibited in the *Rhythmic* one (Delta = Wait Time Contingent – Wait Time Rhythmic), both when the session was easier (i.e., in Italian) or more difficult (in English). In Fig. 6 these differences are plotted for each participant, and participants are ordered as a function of their average writing speed. From the graph it is clear the most subjects (5 of 8) exploited the contingency in the easier condition to reduce the Wait Time (i.e., most blue bars are negative). However, all but the fastest subject showed the opposite tendency in the more difficult (English) task. So there is a significant difference in the strategy and relative timing adopted as a function of task difficulty, even within the same subject (Pair sample t-test on Delta with Language as factor, t(7)= -3.31 p =0.013).

IV. DISCUSSION

The aim of this study was two-fold. On one hand we wanted to demonstrate the importance for the robot to read an implicit communication signal as the establishment of mutual gaze to regulate the interaction. On the other hand we aimed at assessing under which conditions a gaze contingent, personalized response could lead to a more efficient or more pleasant interaction.

To begin with the first question: was the possibility for the robot to monitor humans' gaze signal important to establish a natural and seamless interaction? The answer that comes from our study seems to be positive. Most subjects did not realize that the difference between the two experimental

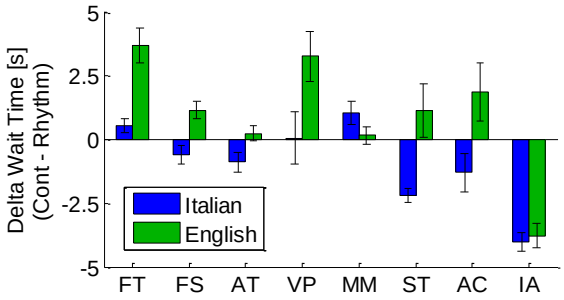

Figure 6: Average differences in Wait Time between the last block of the Contingent and the *Rhythmic* conditions for the different subjects. Error bars represent standard deviations (SD) of the difference.

conditions was the gaze–dependency of one of them. That notwithstanding for most of them the interaction proceeded successfully both in the *Rhythmic* and in the *Contingent* case, suggesting that looking at the robot when the writing was completed came as a quite natural attitude. Only one subject needed an incitement by the experimenter ("Please consider that the robot is waiting for your interaction before continuing the dictation"), but also this subject after this first comment started a natural turn-taking with the robot. Hence, reading subjects' gaze was an efficacious "trick" to appropriately time the robot's actions, leveraging on a natural human attitude, i.e. looking at a silent speaker to get more information.

But was mutual gaze the better signal to indicate readiness in this setting? Alternative approaches could have been monitoring the hand of the writer to be able to anticipate when he was decelerating toward sentence completion. Although such an approach could have guaranteed a higher degree of anticipation in the contingent session and hence a higher responsiveness, it would have increased also the complexity of the system. In particular establishing when the subjects are satisfied with their writing and ready to pass to the next sentence can be an insidious task. Should the right time be at the end of writing of the dictated sentence? And what if the subject feels the need to re-read or correct a misspelled letter or to add punctuation? The use of gaze as implicit signal actually moves the responsibility of the choice of when to pass to the new sentence directly to the subject, who freely and unconsciously decides when he is ready. Moreover, monitoring the gaze of multiple people at the same time is already possible with our system, while monitoring multiple people's writing could represent a more challenging and error prone task. This consideration could become relevant in view of possible applications or robot teaching groups of people, for instance at a school.

Moving to the second main question of our work, establishing whether a contingent behavior is advantageous in a robot also when its role is that of a leader (and potentially

a pace maker) requires a more complex evaluation. From a subjective point of view, the answer seems again positive: no participants preferred the *Rhythmic* condition to the *Contingent* one, and five of the 8 subjects felt that the *Contingent* condition was slightly easier and less error prone.

From a quantitative evaluation of the performance however the reply must be more cautious. Although no significant decreases in performance appeared when subjects were – implicitly – allowed to pace the interaction, neither a significant improvement (e.g., faster task completion) appeared on average. Moreover, a few subjects needed some trials before getting entrained with a steady-state rhythm of interaction, which led to highly variable behaviors at the beginning of the *Contingent* session (e.g. compare first and second block of trials in IA and AT in Fig.4, top and bottom panel respectively). Furthermore, choosing a contingent approach implies also the increase in the risk of system errors, that a simpler rhythmic system does not face. So, a trade-off must be evaluated between the advantages yielded by the contingency and the inherent risks of errors (in our case false positives or false negatives in the detection of subject's mutual gaze – not detecting the readiness of the subject, see Sec. III A for a quantification in our settings).

On the other hand, the subjects who were at the extremes of the writing speed distribution – the slowest and the fastest – could actually take advantage the robot contingent behavior. Only in such condition the former could complete writing without being interrupted by the next dictation and the latter could accelerate the dictation process at her own pace. So, if the *Rhythmic* condition is good enough for the average subject, the *Contingent* case makes a real difference mostly for the outliers. This trend is visible also within the same subject when faced with tasks of different difficulty. Indeed, most participants exhibited the opposite strategy when dealing with an easier or a more complex task: they decreased their Wait Time when writing in their mother tongue and increased it for the dictation in the foreign language (see Fig.6). So, a contingent approach makes the robot dictation suitable to cope with variability among different subjects and also within the same subject, if dealing with tasks characterized by different levels of difficulty.

To sum up, although a contingent behavior in our dictation context had clear subjective advantages and did not disrupt the appropriate rhythm of the interaction, a case to case evaluation is required to quantify the advantages and drawbacks that such an approach might determine. Indeed, contingency might lead also to the need for an initial adaptation to the turn-taking and to a larger variability in subjects' performances as a function of task difficulty. However, if the implementation of a contingent system is sufficiently simple and robust, there are situations in which it should be preferred. In particular this holds true when an average estimate of human behavior is not a good predictor for the performance of the individual human partner involved in the interaction, as for instance when working with kids or special populations).

V. Conclusion

"Taking dictation requires choreography between speaker and listener" [18] and such a choreography can be achieved through a rhythmic leading of the teacher or through an adaptive, gaze-contingent interaction between speaker and listener. We have shown that this latter approach makes the interaction more comfortable and consequently preferable to the majority of the subjects. However, the larger quantitative benefits are not for all of them but rather for the outliers, as the contingent approach allows them to exploit (or cope with) their specific characteristics. Therefore, a principle as simple as detecting the establishment of mutual gaze becomes for a robot an efficient mean to seamlessly interact with human partners with different needs in a turn-taking task.

References

[1] N. George and L. Conty, "Facing the gaze of others.," *Neurophysiol. Clin.*, vol. 38, no. 3, pp. 197–207, Jun. 2008.

[2] B. Mutlu, T. Shiwa, T. Kanda, H. Ishiguro, and N. Hagita, "Footing In Human-Robot Conversations : How Robots Might Shape Participant Roles Using Gaze Cues," vol. 2, no. 1.

[3] J. G. Trafton, M. D. Bugajska, B. R. Fransen, and R. M. Ratwani, "Integrating Vision and Audition within a Cognitive Architecture to Track Conversations," in *ACM/IEEE International Conference on Human-Robot Interaction*, 2008.

[4] K. S. Lohan, K. J. Rohlfing, K. Pitsch, J. Saunders, H. Lehmann, C. L. Nehaniv, K. Fischer, and B. Wrede, "Tutor Spotter: Proposing a Feature Set and Evaluating It in a Robotic System," *Int. J. Soc. Robot.*, vol. 4, no. 2, pp. 131–146, Dec. 2011.

[5] S. Andrist, X. Z. Tan, M. Gleicher, and B. Mutlu, "Conversational gaze aversion for humanlike robots," *Proc. 2014 ACM/IEEE Int. Conf. Human-robot Interact. - HRI '14*, pp. 25–32, 2014.

[6] D. G. Novick, B. Hansen, and K. Ward, "Coordinating turn-taking with gaze," *Proceeding Fourth Int. Conf. Spok. Lang. Process. ICSLP '96*, vol. 3, 1996.

[7] G. Metta, G. Sandini, D. Vernon, L. Natale, and F. Nori, "The iCub humanoid robot : an open platform for research in embodied cognition.," in *Proceedings of the 8th Workshop on Performance Metrics for Intelligent Systems*, 2008, pp. 50–56.

[8] U. Pattacini, "Modular Cartesian Controllers for Humanoid Robots: Design and Implementation on the iCub," *PhD Dissertation*, 2011.

[9] A. Sciutti, L. Schillingmann, O. Palinko, Y. Nagai, "A Gaze-contingent Dictating Robot to Study Turn-taking," in *Proceedings of the 10th ACM/IEEE International Conference on Human-Robot Interaction*, 2015.

[10] D. E. King, "Dlib-ml : A Machine Learning Toolkit," *J. Mach. Learn. Res.*, vol. 10, pp. 1755–1758, 2009.

[11] V. Kazemi and S. Josephine, "One Millisecond Face Alignment with an Ensemble of Regression Trees," in *Computer Vision and Pattern Recognition (CVPR), 2014*, 2014.

[12] F. Timm and E. Barth, "Accurate eye centre localisation by means of gradients," in *International Conference on Computer Theory and Applications ({VISAPP})*, 2011, pp. 125–130.

[13] B. A. Smith, S. K. Feiner, and S. K. Nayar, "Gaze Locking : Passive Eye Contact Detection for," *UIST*, pp. 271–280, 2013.

[14] M. Schröder and J. Trouvain, "The German Text-to-Speech Synthesis System MARY: A Tool for Research, Development and Teaching," in *International Journal of Speech Technology*, 2003, vol. 6, pp. 365–377.

[15] C. M. Brown, "Human-computer interface design guidelines," *Ablex Publishing Corp.*, Jan. 1988.

[16] H. Brugman and A. Russel, "Annotating multi-media/multi-modal resources with ELAN," *Int. Conf. Lang. Resour. Eval.*, pp. 2065–2068, 2004.

[17] M. Hanheide, M. Lohse, and A. Dierker, "SALEM - Statistical AnaLysis of Elan files in Matlab," in *Multimodal Corpora: Advances in Capturing, Coding and Analyzing Multimodality*, 2010, pp. 121–123.

[18] K. Johnson and E. Street, "Response to intervention and precision teaching: creating synergy in the classroom," 2011.