Abstract—We study the problem of handling the inter-turn pauses in a human-robot dialogue. In order to reduce the impression of elapsed time while the robot transcribes, understands and starts uttering a response we propose to automatically generate conversational fillers, to fill the silences. These fillers combine verbal utterances with body movements. We propose a Bayesian model that samples filler whose production duration time is close to the expected computational time needed by the robot. To increase the sensation of engagement, the fillers also include contextual information gathered during the dialogue (such as the name of the interlocutor), if this information is present with high confidence.

We evaluate this approach with an indirect user study measuring time perception, comparing three different strategies to overcome the inter-turn time (silence, static filler and our approach). The results show that users prefer the dynamic fillers, even when the conversation is objectively shorter with one of the other strategies.

I. MOTIVATION

Smooth human-machine conversations are a promise of the near future. Recent advances in natural language technologies and infrastructure make it seem feasible that current one-shot queries (like Google Search, Siri, Alexa, etc) will soon take the form of true dialogues. It seems therefore natural to extend such conversations to social robots, robots that are meant to socially interact with humans (as opposed to industrial robots). However, the newest such technologies which can be found on the market are far away from that promise. The main reason for that is that they do not perform full-fledged automatic speech recognition (ASR) but only look for a limited number of expected keywords, which have to be predefined. Scenario designers then have to “think of any possible word that a user could utter in this context”

This happens not because full ASR systems are not available or bad: progress in neural networks has spurred impressive improvements in this domain in the last years, and several companies provide access to their ASR services through a cloud offering. Latency however, is a problem. While relatively fast, they still require non-negligible time to upload the speech and recognize it (in our benchmarks this was around 0.4 s wrt the length of the utterance). Although incremental ASR techniques promise to reduce this problem, availability of such systems is extremely reduced and current commercial offerings (which have the necessary data to provide state-of-the-art results) are mostly batch-based.

As reported previously [1], in human-robot conversations, humans prefer to receive an answer between 1 and 2 seconds after they finished their utterance, which is lower than what it usually takes the system to understand, react and start uttering an answer. That expectation from the human side has probably a relationship to implicit models that humans use when they model expected duration of inter-turn silences. When the actual duration is different (either shorter or longer) than that expectation, it lead to a lack of trust in the answer [2]. Longer pauses in particular create problems due to a lack of engagement. In general, passivity and low information load leads to boredom [3] which itself produces low satisfaction [4], adding to other problems to achieve smooth human-robot interactions. Citing Zakay [5]: “when information processing load is below an optimal level for a specific individual a feeling of boredom is raised. Boredom is accompanied by a slowing of the felt pace of the flow of time.”

We propose to overcome these silences with conversational fillers, both verbal and non-verbal. While several studies have acknowledged the importance of handling the pauses in a human-robot dialogue, we are aware of very few proposals on how to overcome them. Moreover, they are overly simplistic (for example: just emitting a periodic sound signaling “still working”).

In this paper, we propose a system that selects multimodal fillers according to a variety of factors, including the expected length of the silence and the context of the dialogue so far. We assume the state of the dialogue to be latent, and propose a Bayesian derivation that includes it directly when computing the probability of selecting a filler.

II. RELATED WORK

While research in dialogue systems is mostly concerned with the content of the dialogue, in this paper we focus on the secondary signals that make a robot-human conversation truly interactive. It is well established that those secondary signals fill out a variety of tasks in a dialogue, including signaling the intention to hold the turn, take it or act as a manifestation of engagement [6]. In this sense, this is related to the concept of backchanneling (defined as “a sound or gesture made to give continuity to a conversation by a person who is listening to another”), and the associated challenge of detecting [7], [8] or synthesizing [9] them. However, our main concern is to keep the user engaged in the dialogue, in particular during the pauses in which the robotic system is processing information.

In our transactional-driven scenario of human-robot dialogue, holding or taking a turn is less crucial than, for instance, in multi-agent conversations [10]. At the same time, the former is the scenario where most business-interest seems

1 dixit from a company deploying social robots in retail
to be focused on lately. While still important, changing the floor-holder is less ambiguous, and those fillers serve mostly the purpose of keeping the engagement [11] and signaling understanding [1].

Probably the most typical conversational fillers in English are uh and um. Clark and Tree [12] summarize 16 interpretations of them, where the most popular ones are “used to announce the initiation of what is expected to be a minor, or major, delay in speaking”. The same article also measures correlation of the length of the following pause with the fact that the pause was preceded by one of these fillers. Similarly, disfluencies have been studied [13], with Betz and Wagner [14] explicitly calling out their potential for “buying time” in speech synthesis.

Several studies on acceptance of human-machine dialogue have stressed the importance of managing well the pauses. Ohta et al. [15] study the effect of silences and filled pauses in the perceived naturalness and the comprehension of transmitted facts. They conclude that adding such pauses helps, in line with Shiwa et al. [1] who report that answers with less than 1s delay are actually worse perceived than those in the 1-2 second window (with perceived naturalness decaying after that). It should be noted that these studies were performed in Japan, and that some research suggests that Eastern cultures accept silence better than Western cultures [16].

But the other extreme is of course also true: the fact that long pauses are a major problem of acceptance and naturalness of human-robotic interaction has long been reported. This holds true, even when the robot is teleoperated [17]. Ohshima et al. [18] studied the impact of adding both body and verbal fillers in human-robot dialogues on how “awkward” the dialogue was perceived, concluding that “when, during a conversation with a robot, it is the human’s turn to speak following an awkward silence, the robot’s fillers enhanced its ability to express a cooperative attitude in order to recommence the conversation”. Similar to an earlier study of the same group on 3D simulations [19], they focus on the impact of those fillers on the dialogue without further exploration on how and when to emit them. Our work is however strongly inspired by their conclusion that “if robots behave appropriately by inserting verbal/bodily fillers during silences because of this time consuming process, humans might more easily understand their situations and estimate/assume that it is going to begin speaking”.

Although text and sounds remain the main channel of communication, conversation with a robot allows for a much richer interaction than only an exchange of sounds. Besides body movements, Andrist et al. [20] studied gaze aversion to hold the floor during the conversation, built on top of studies that argue that such behaviour signals cognitive effort.

We have found very few work on how to add these fillers during a conversation. Bohus and Horvitz [11] mention that hesitation mechanisms can be used to “buy more time”, but the focus there is on avoiding disengagement only, and the hesitation is triggered as soon as a low level of engagement is reached. Skantze et al. [10] propose a system for multi-party conversation with a real-time component that decides to produce a turn-taking clue (filled pause, gaze aversion or inhalation) or a backchannel to pass the turn. One of the few studies we are aware of that proposes to generate conversational fillers in dyadic multi-turn dialogues is Shiwa et al. [1]. While they report positive results by doing so, the technique they use is overly simplistic: they just repeat etto (“I am thinking”) until the computation is completed. It is therefore a direct adaptation of the hourglass cursor and similar signals indicating that a computer is processing. Glaser et al. [17] extend this in the context of a teleoperated robot: they measure the time it takes the human operator to fulfill a task, and produce conversational filler (verbal ones only) if this estimation exceeds two seconds. For time periods that are expected to be short (less than seven seconds) they just utter etto, while they use longer ones for longer duration. The decision how to fill a pause is, thus, made only on the basis of a task’s type, does not take into account the input utterance and focuses on the expected time of the human operator.

An obvious question is if our proposal makes sense considering the advances made in incremental dialogue systems in the last years [21], [22], [10]. Such frameworks however are still far from trivial to set up, and are not compatible with existing cloud-based providers of ASR or NLU which have shown a dramatic increase in popularity in the last year. Moreover, our formulation of the problem under a Bayesian lens allows to include other information as priors, and can therefore be extended.

III. PROPOSAL

We will formalize the dialogue as a list of utterances, spoken alternately by the robot and the human. A started dialogue will be denoted by \( d_k = [r_1, h_1, r_2, h_2, \ldots, r_k] \), where \( r_i \) stand for utterances spoken by the robot and \( h_i \) for utterances spoken by the human. Furthermore, we assume that such a dialogue can be summarized by a state \( s(d_k) \) which encompasses the relevant information so far. For transaction-oriented dialogues, this includes at least the intent of the human (what kind of transaction the user wants to perform) and the information gathered so far (commonly called slots). It can also include other signals such as a detected emotion and personality (of the robot or the person), inferred engagement [11], etc. It is the job of the dialogue manager to keep a distribution of the state of the dialogue \( p(s(d_k)|d_k) \), and we assume this as given here. To ease the notation, we will drop the subscript \( k \), as it will be clear from the context (we will only consider the distributions at a fixed time \( k \)).

At the moment when a human finished its utterance \( h_i \), the different modules of the dialogue system start to perform four actions in sequence:

1) transcribe the speech (automatic speech recognition)
2) understand it (natural language understanding (NLU))
3) decide on the next action (dialogue manager)

Four actions in sequence:
4) start uttering a natural language realization of that action (natural language generation)

We will call the aggregate time to execute these four actions **time-to-process**, denoted by $t(s(d), h)$.

Our goal then is to produce a conversational filler that takes time close to $t(s(d), h)$, given the unfinished dialogue $d$. Of course $h$ itself is unknown, as transcribing it may be the bulk of the **time-to-process**. However, we can easily compute the size of the finished utterance and try to predict the **time-to-process** knowing $s(d)$ and $len(h)$ (the length of $h$); which we will denote by $t'(s(d), len(h))$. $t'$ was estimated offline, using past corpus.

Our main goal then is to estimate a probability distribution of filler actions $a$, given the dialogue $d$, the length of the sound envelope $len(h)$ and parameters of the robot $\theta$, which controls global dialogue parameters (such as personality, formality of the dialogue, etc). This distribution should be correlated with the **time-to-process**, in the sense that the time to execute the filler action $a$ (denoted by $t_a(a, s(d))$ should be close to $t'(s(d), len(h))$. In formula, we want to estimate $p(a|d, len(h), \theta)$. To compute this, we will use the latent variable $s(d)$ as follows:

$$p(a|d, len(h), \theta) = \sum_{s(d)} p(a|s(d), len(h), \theta)p(s(d)|d, \theta) \quad (1)$$

which comes from a direct application of Bayes rule, the fact that $s(d)$ summarizes the dialogue $d$ and that $s(d)$ is independent of the last utterance $h$. In this formulation, the production of filler is tightly linked to the overall dialogue manager as it includes the information present in the dialogue state.

IV. METHODOLOGY

We implemented this idea on top of a humanoid robot, using an in-house virtual agent that performs transactional queries. The use-case is that of a reception, where the virtual agent first tries to understand the reason of a visitor’s arrival (classifying it into one of several predefined categories such as meeting an employee, deposing a parcel, asking for information, etc), and then gathers the necessary information to process the transaction or fulfill the information need. The selected platform is NAO, a 58 cm tall humanoid robot produced by SoftBank Robotics Europe, which we interfaced with the ASR system of a major software company.

We created several verbal conversational fillers, and combined them with different body movements including full-body movements, arm, head and the different LED’s the robot has. Extensions towards other mechanisms (such as a screen or non-voice sounds) are of course possible. An example list of conversational fillers is available in Table I. For the body movements we noted the time it took them to complete, and for the verbal fillers we assumed a speech rate of 150 words-per-minute, the average rate for English speakers according to the National Center for Voice and Speech.

This allows us to automatically compute the time $t_a(a, s(d))$ and accelerates the creation of new fillers. Note how we combine the slots retrieved so far by the dialogue manager (such as NAME) in order to provide a more contextualized feedback during the conversations (eg: Please wait a second, {NAME}). This also influences the computation of $t_a$, the time it takes the specific filler to be uttered.

For the **time-to-process** $t'(s(d), len(h))$, we estimate it with a linear function $c \times len(h) + b$, which is independent of $s(d)$. To estimate the values of $c$ and $b$ we run several experiments and fixed $c$ to be 0.4 and $b$ to 1 (second), which accounts for a fixed overhead necessary to communicate with the server hosting the dialogue manager. That the **time-to-process** is linearly dependent with the length of the spoken utterance is specific to this use-case, where the decisions are simple to take and almost all of the time is consumed by the ASR module. If the decision would require more sophisticated treatment (access to another knowledge-base for instance) this should be taken into account.

In order to make the computation of Eq. 1 tractable, we made some simplifying assumption. We set $p(s(d)|d)$ to 0, for small values: such a non-smooth estimation means that we only have to account for states for which there is a high probability during the dialogue. Furthermore, for filler actions that require certain slots to be filled (in our examples in Table I these are name and affiliation) we fix $p(a|s(d)) = 0$ if the corresponding slots in $s(d)$ are not filled (or if they are with a low probability). Besides that we do not include any other preference of one filler over another. As a consequence of this, we made sure that fillers were generic enough to be used in a variety of contexts. Learning when to utter a certain filler, based on the content of that filler (besides its length) is beyond the scope of this paper.

Finally, we assume that $p(a|len(h))$ is proportional to $|t'(s(d), len(h)) - t_a(a, s(d))|^{-1}$, which, after replacing $t'$, gives:

$$p(a|len(h)) \propto |c \times len(h) + b - t_a(a, s(d))|^{-1}$$

$\theta$ can then be used to give more weights to more generic fillers or down-weight humorous ones. Assume – as a simple example –, a potential filler “Ok {}, gimme a second, slots=['name']”. In order to get selected the state $s(d)$ should contain a slot for the variable $name$; $t_a(a, s(d)) = 2s$ as it only contains words and no modifiers or body movements (5 * 60/150) and its probability to be selected would be proportional to $|0.4 \times len(h) - 1|^{-1}$ and would therefore be maximal for an utterance that took 2.5s.

V. EVALUATION

To appreciate the values of these fillers, and the impact it has on the conversation we recorded an interaction with and without these fillers. The dialogues are not exactly the same (due to randomness associated in the generation part) although the semantics of the interaction is equivalent. They follow a typical transactional framework of three stages: intent detection (detecting what the visitor wants),

\[\text{http://www.ncvs.org/ncvs/tutorials/voiceprod/tutorial/quality.html}\]
slot filling (depending on the intent, retrieve a pre-defined set of information) and fulfillment (final action, which in this case was just a request for final confirmation). A complete interaction can be seen at http://videos.xrce.xerox.com/index.php/videos/index/875: these is slightly more elaborate than the video shown to the evaluators (which basically stopped before placing the phone call).

There are a total of 5 turns for a final duration of about 120 seconds. It is longer when using conversational fillers, mainly from non-optimized procedures that run on the limited CPU of the robot. We performed some preliminary study with users having or observing the conversations and asking direct questions about which filling strategy they thought was more engaging, less annoying or less boring. When asked such direct questions, the answers favored overwhelmingly our dynamic approach. Therefore we set out to explore if we could see an impact on an indirect and more challenging (but also more valuable) measure.

The ultimate goal is to engage the human more deeply into the interaction. Using fillers does not reduce the overall interaction time (actually, it may even increase it), but there is a vast literature (as well as common sense) that points towards a strong correlation between boredom and time perception. Time-perception increases with boredom [3], [5], which itself has important links with satisfaction and sensation of fulfillment [4] but can be reduced by providing mental stimulus in the form of novelty. Following the conclusions from these psychological studies, we decided to measure time-perception as a proxy for engagement. In this sense, instead of performing a direct evaluation which would have favored clearly our approach, we performed an indirect one: users were asked to choose which conversation felt shorter. Because individual conversation may last longer (even for the same scenario) due to variable-length utterances of the human and the robot (the natural language generation engine samples from a variety of surface forms) we recorded conversations with all three strategies for two different scenarios. It can be argued that the time perception of observing a dialogue is different than the one when participating actively in it. However, an active dialogue has several free parameters which are hard to account for and would require substantially more data-points to extract meaningful results.

We compared our approach against the obvious baseline of not using any filler, but also against the strategy of using static fillers which simply uttered the same filler once per turn (hum-hum in our case).

### A. Setting and Guidelines

Volunteers from our institution were then requested to watch those recordings and answer the question “Which of these videos felt shorter?”, and had the option of adding free-text comments. They were instructed to not time the duration, and any indication of the time (progress bars or display of time spent/remaining) was removed from the video player. We requested prospective duration instead of retrospective: the users in our study know from the beginning that they will have to evaluate the perceived time elapsed. Previous studies confirm that the variance in prospective duration judgments is smaller than in retrospective [23].

We recorded the scenarios a priori, and evaluators watched them through a web-service. The scenario was sampled randomly, and the users only compared two randomly displayed videos for each scenario (users could annotate more than once, which a few did). A snapshot of the interface
can be seen in Fig. 1. We note that the No scenario was either equally long or 4 to 10 seconds shorter than their filling counterparts (all videos last approximately one and half minute). We asked voluntary participation through an institution-wide distribution list, and participants received an ice-cream. Only three participants were familiar with our general research project, and none was disclosed about the specific goal of this evaluation.

In total, 32 different annotators (42% female, average/median age 38/36) provided 64 annotations.

VI. RESULTS

Table II shows the result of these pair-wise comparison. We measured statistical significance through a binomial one-sided test (testing the hypothesis that the percentage from the second column is greater than random). When comparing the No-filler approach against the static version, the static filler version was slightly preferred. The statistical test however shows that such an approach has 1 in 3 chances of being a random phenomena. As one annotator put it, the static filler “emphasizes the time when nothing happens”. For the dynamic fillers the preference was more convincing, with two-third of the users preferring the dynamic version over the silent one; with a convincing statistical support (significant with $p<0.1$). The static vs dynamic comparison also favors the dynamic approach, but the statistical test provides only a weak support for that. However, it should be noted once more that the static filler version is objectively shorter than the dynamic one (10 seconds for a total duration of 120 seconds)$^3$.

The comments go in the same lines, with several of the users pointing out that in the static filler scenario: “hum hum is extremely annoying”, and that the long pauses of not using any filler was “nerve-racking”. On the other side, several commented that the dynamic version felt more “fluid”.

VII. CONCLUSIONS

We present a novel algorithm for context-aware selection of conversational fillers in human-robot dialogues. The importance of handling the silences in a conversation is well established in the literature, and we build upon these previous studies, proposing a novel method that selects the right utterance based on the current context of the dialogue. As an example, we created a list of verbal conversational fillers and gestures accompanying them and integrated them into a dialogue system built on top of a NAO humanoid robot. This list can be easily expanded and adjusted for other robotic platforms.

Unlike other techniques dealing with time delays, our system takes into account the context of the dialogue and selects the filler that has the right duration, and includes information gathered so far in the dialogue. The problem is interesting because the filler can not use the content of the immediately previous utterance, as the time to transcribe it is in general the bulk of the inter-turn silence. For this we rely on the envelope of the sound, and produce fillers that are consistent with the length of the previous utterance adding randomness for the purpose of variance.

By including multi-modal filled pauses into the dialogue, we are able to fake robot’s signals of understanding and make additional steps towards a natural interaction between the user and the robot. Our user study in this sense showed a clear preference towards such dynamic fillers, and that the quality of human-robot interaction benefits from context-aware controlling of time delays. When comparing two equivalent dialogues, the users preferred this approach of dynamic fillers, even when the other video was objectively shorter.

While we present our filler as separate module that can be used in a dialogue systems, its Bayesian formulation allows to include internal information (probability of the current state for instance) as a prior.

REFERENCES


<table>
<thead>
<tr>
<th>Comparison</th>
<th>user impression (which felt shorter)</th>
<th>p-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>No vs Static</td>
<td>66.67% static</td>
<td>0.094</td>
</tr>
<tr>
<td>No vs Dynamic</td>
<td>60.00% dynamic</td>
<td>0.251</td>
</tr>
<tr>
<td>Static vs Dynamic</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

TABLE II

QUANTITATIVE RESULTS OF OUR USER STUDY

$^3$Another way of showing this is that if we would like to test the hypothesis that the static version is shorter (which it is), based on the human judgement, the associated $p$ value would be of 0.808.
Please watch and listen both videos, and then answer the questions below.

Video A

Video B

Questions:
1. Which of these videos felt shorter? Video A Video B

Fig. 1. Snapshot of the annotation interface.