Teaching Computers to Conduct Spoken Interviews: Breaking the Realtime Barrier with Learning

Gudny Ragna Jonsdottir and Kristinn R. Thórisson

Center for Analysis & Design of Intelligent Agents and School of Computer Science
Reykjavik University
Kringlunni 1, IS-103 Reykjavik, Iceland
{gudny04,thorisson}@ru.is

Abstract. Several challenges remain in the effort to build software capable of conducting realtime dialogue with people. Part of the problem has been a lack of realtime flexibility, especially with regards to turntaking. We have built a system that can adapt its turntaking behavior in natural dialogue, learning to minimize unwanted interruptions and "awkward silences". The system learns this dynamically during the interaction in less than 30 turns, without special training sessions. Here we describe the system and its performance when interacting with people in the role of an interviewer. A prior evaluation of the system included 10 interactions with a single artificial agent (a non-learning version of itself); the new data consists of 10 interaction sessions with 10 different humans. Results show performance to be close to a human's in natural, polite dialogue, with 20% of the turn transitions taking place in under 300 msecs and 60% under 500 msecs. The system works in real-world settings, achieving robust learning in spite of noisy data. The modularity of the architecture gives it significant potential for extensions beyond the interview scenario described here.

Keywords: Dialogue, Realtime, Turntaking, Human-Computer Interaction, Natural Communication, Machine Learning, Prosody.

1 Introduction

One of the challenges in giving computers the ability to participate in spoken dialogue is getting them to perform such activity at a natural pace. Although people can to some extent adapt to the often stilted interaction resulting from a system's lack of human-like turntaking, a system that adapts to human speaking style would be vastly preferable to one which requires its users to change their natural speaking style. In this paper we describe our work on building a flexible dialogue system, one that can adapt in realtime to a person's speaking style, based on prosodic features. As a framework for testing our theories we have created an artificial agent, *Askur*, that uses prosodic features to learn "polite" turntaking behaviors: minimizing silences and speech overlaps. Askur learns this on the fly, in natural, full-duplex (open-mic) dynamic interaction with humans.

Zs. Ruttkay et al. (Eds.): IVA 2009, LNAI 5773, pp. 446–459, 2009.

In natural interaction mid-sentence pauses are a frequent occurrence. Humans have little difficulty in recognizing these from proper end-of-utterance silences, and use these to reliably determine the time at which it is appropriate to take turn – even on the phone with no visual information. Temporal analysis of conversational behaviors in human discourse shows that turn transitions in natural conversation take on average 0-250 msecs [1,2,3] in face-to-face conversation. Silences in telephone conversations - when visual cues are not available - are at least 100 msecs longer on average [4]. In a study by Wilson and Wilson [1] response time is measured in a face-to-face scenario where both parties always had something to say. They found that 30% of between-speaker silences (turn-transitions) were shorter than 200 msecs and 70% shorter than 500 msecs. Within-turn silences, that is, silences where the same person speaks before and after the silence, are on average around 200 msecs but can be as long as 1 second, which has been reported to be the average "silence tolerance" for American-English speakers [5] (these are thus likely to be interpreted by a listener as a "turn-giving signal"). Tolerance for silences in dialogue varies greatly between individuals, ethnic groups and situations; participants in a political debate exhibit a considerably shorter silence tolerance than people in casual conversation – this can further be impacted by social norms (e.g. relationship of the conversants), information inferable from the interaction (type of conversation, semantics, etc.) and internal information (e.g. mood, sense of urgency, etc.). To be on par with humans in turntaking efficiency a system thus needs to be able to categorize these silences.

Artificial agents that can mactch humans in realtime turntaking behavior have been slow in coming. Part of this is due to poor collection of realtime behavioral data from interlocutors. A vast majority of current speech recognizers, for example, use silence detection as the *only* means for deciding when to reactively start interpreting the preceding speech. This leads to unnatural pauses, often one, two, or even three seconds in length, which may be acceptable for dictation but is ill-suited for realtime dialogue. Part of the challenge, therefore, is to get the system to behave quickly enough to match human-style interaction. However, achieving such low-latency turn transitions reliably cannot be done reactively [6]; to have any hope of achieving the 200-500 msec levels observed in human dialogue requires the system to *predict* what actions to take. This must be done using realtime perceptual data collected of interlocutor (unimodal or multimodal) behavior. As inter-subject and real-world scenario complexity puts practical limitations on the amount of hand-coding that can be brought to bear on the problem, the most sensible way to approach this problem is to engineer the system to automatically learn which features of speech can be used for this purpose.

We want to build a general learning mechanism that can automatically learn complex turntaking cues in realtime dialogue with human users. Our approach is based on the Ymir Turntaking Model (YTTM), which models turntaking as a negotiation process controlled jointly by the participants through loosely coupled perception-cognition-action loops [6], and proposes modular construction blocks for this purpose. The original implementation of this model has been expanded

according to the Constructionist Design Methodology principle [7], to incorporate learning mechanisms. These allow the system to adjust to interlocutors in realtime and learn over time, achieving human-like performance characteristics in under 30 turns. To the best of our knowledge no system has so far been described in the literature that can adjust its turntaking style dynamically to individuals to achieve human-like performance characteristics, while continuing to improve its performance as it interacts with more people.

In Jonsdottir and Thórisson (2008) [8] we described the first version of the system and presented data on its learning ability when interacting with another artificial agent (a non-learning copy of itself), listening for features of the prosody of the Loquendo speech synthesizer to determine its turntaking predictions and behavior. The results, while promising, described interaction sessions between the system and a single synthesized voice, with negligible noise in the audio channel. Even though the learning in such a controlled artificial setup proved successful we did not consider this to be a guarantee that it would generalize to a real-life setting when conducting live interviews with people. To evaluate the learning mechanism in a more realistic scenario we configured the system to conduct realtime interviews with people over Skype. The interviews were designed to require no natural language processing[1], only prosodical features inform the behavior of the system as it learns to minimize its silences while trying to avoid overlaps.

After reviewing related work we describe the architecture of the learning system, the experimental setup, and then the results of the human subject study, showing how the system learns during the interaction.

2 Related Work

Sacks et al. [9] and Walker [10] were among the first to point out the possible role of prosody and intonation in enabling people to take smooth turns. Walker made a well-informed argument that conversants employ *relatime processing of prosodic information contained in the final few syllables of utterances* to determine when the appropriate moment is to give back-channel feedback, as well as take turn.

J.Jr. was an early computer agent demonstrating this ability [11]. Using realtime processing of a person's prosody, the system could analyze it fast and accurately enough to interject back-channel feedback and take turns in a highly human-like manner. The subsequent Gandalf agent [12] adopted key findings from J.Jr., based on the Ymir framework, an expandable granular AI architecture. Gandalf analyzed in realtime an interlocutor's gaze, gesture, body stance and prosody to determine appropriate turntaking and back-channel opportunities. This has been done more recently in the Rapport Agent [13], which uses

[1] The exclusion of speech recognition and language interpretation in this paper is a limitation of the current research focus, not a general limitation of the dialogue architecture we are developing.

gaze, posture and prosodic perception to, among other things, detect backchannel opportunities. While performing in realtime, approaching human-level pace in some cases, none of these systems were built to adapt their behavior to their interlocutors.

Although Reinforcement Learning and other learning methods have been used to some extent in dialogue systems, most of these attempts have been done via offline training of the system. Sato et. al [14] use a decision tree to enable a system learn when a silence signals a wish to give turn and Schlangen [15] has successfully used machine learning to categorize prosodic features from a corpus. Morency et al. [16] use Hidden Markov Model to learn feature selection for predicting back-channel feedback opportunities. However, by these studies, by and large, ignore the *active* element in dialogue – the need to test the quality of perceptual categorization by generating realtime behavior based on these, and monitoring the result. As dialogue is a realtime negotiation process [17] any such effort must include both parties in interaction in order to generalize to real-world situations. (A negotiation process of two parties cannot be simulated without including the effect that the behavior of one has on the other – in a realtime feedback loop.) Classifying perceptual features is certainly one step, but doing so in realtime is another, and generating behaviors based on these – behaviors that affect the other party in some way – yet a third one.

The Ymir Turntaking Model (YTTM, [6]) addresses realtime multimodal turntaking, taking both perception and action into account. YTTM specifies how perceptual data are integrated to derive *how and when* certain perceptual, turntaking and dialogue acts are appropriate, and how to behave according to such information. While the YTTM does not address learning it is based around a modular formalism that has enabled us to add such capabilities without restructuring its original model. We have based our approach and system on this model.

3 System Architecture

The goal of our work is to create a dialogue system that interacts at human speed and accuracy in natural conversation. With primary focus on incremental processing, adaptability and error recovery, our system autonomously learns to predict appropriate turntaking behaviors so as to minimize both awkward silences and overlapping speech in two-party realtime conversation with a human. Our speaking agent, Askur, performs this task by learning to appropriately adjust its silence tolerance during the dialogue (See Figure 1).

The architectural framework is described in more detail in [8] and [18]; a quick review of this work will aid in understanding what follows. The architecture, which is in continuous development, currently consists of 35 interacting modules in a publish-subscribe message passing framework. Its modularity and separation of topic knowledge and behavioral knowledge make it relatively easy to install and test specific "communication skill" components within the framework, compared to alternative approaches. The Ymir Turntaking Model

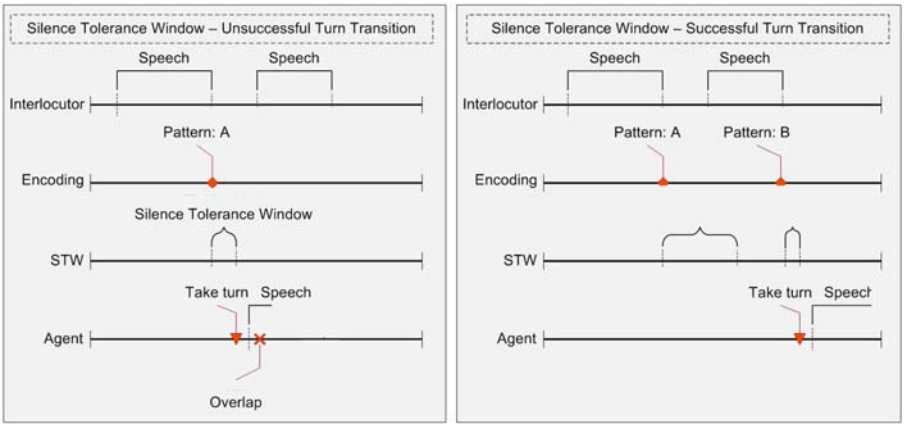

Fig. 1. The interlocutor's speech is analyzed in realtime; as soon as a silence is detected the prosody preceding the silence is decoded. The system makes a prediction by selecting a Silence Tolerance Window (STW), based on the prosody pattern perceived in the interlocutor. This window is a prediction of the shortest safe duration to wait before taking turn: a window that is too short will probably result in overlapping speech while a window that is too large may cause unnecessary/unwanted silence.

(YTTM) provides the backbone of the system [6]. Multi-modal deciders use information extracted from the external (perceptual) and internal environment, to synchronize the perceived and anticipated contexts, which in turn steer both perceptual and behavioral system events. Perception modules include speech and prosody analysis. Prosody analysis is generated as a stream, with an 80 msec fixed latency; speech interpretation can easily take up to a couple of seconds[2]. System responses are generated incrementally in a so-called content generation module where the topic knowledge lies. Speech is realized using the Loquendo text-to-speech synthesizer, which currently imparts a 200 msec latency from the decision to speak to the point when the sound of the first syllable of the first word reaches the audio speaker[3]. One way to compensate for this latency is to pre-start execution with the possibility of canceling it gracefully within 100 msecs, or the chosen Silence Tolerance Window (STW, see 1).

Our agent contains a learning framework that is separated from the decision-making through a service interface, with the benefit of thinner deciders and improved reusability of the learning functionality. The mechanism learns correlations between states and actions and is implemented with ϵ-greedy Q-learning algorithm. ϵ-greedy algorithms try to avoid getting stuck in a local maximum by

[2] While speech recognition does not matter to the present topic of learning turntaking, the system is built to include the full spectrum of speech events. Thus, speech recognition is an integral part of the system's architecture.

[3] 200 msecs is long in the context of human-like turntaking skills, but it is by far the best performance of any speech synthesizer we have achieved to date.

exploring less favorable actions in a certain percentage of trials. Q-learning was chosen partly because of this feature and partly because it is model-free, so the agent can start with no knowledge of possible states and actions. Modules that use the learning service contain their own action selection strategy and a general policy on what to do in unseen circumstances. This leaves action discovery solely in the hands of the service-recipient module. The recipient module encapsulates the state-action pair it wants evaluated into a decision for the learner to evaluate. The learner builds a policy of estimated returns for each state-action pair, which it publishes regularly. Each decider that wants to use learning information so published needs to reciprocally publish every decision it makes. A decision contains a state-action pair to be evaluated and a lifetime during which rewards can be assigned; rewards can also be assigned on timeout, representing scenarios where the lack of consequence should be rewarded.

3.1 Feature Selection and Extraction

Prior research has shown that the final part of speech preceding a silence can contain prosodic cues relevant to pragmatics [19]. Following [20] we use the last 300 msecs of speech preceding each silence. The incoming audio signal is handled in a 2-step process. We use the Prosodica prosody analyzer [21] to compute speech signal levels and speech activity. It analyzes prosody in steps of 16 msecs, producing a continuous stream of data from which high level features can be extracted.

Two distinct features are used to categorize all silences. The 300 msecs of the most recent tail of speech right before a silence is searched for the most recent local minimum/maximum pitch to identify the *starting point of the final slope*. Slope is split into three semantic categories: *Up*, *Straight* and *Down* according to formula 1; end-point is split into three groups for *relative value of pitch right before silence*: *Above*, *At* and *Below* the average pitch for the speaker (for the whole dialogue period), according to formula 2. This gives us 9 different combinations of features.

$$m = \frac{\Delta pitch}{\Delta msecs}, \begin{cases} \text{if } m > 0.05 \rightarrow slope = Up \\ \text{if } (-0.05 \leq m \leq 0.05) \rightarrow slope = Straight \\ \text{if } m < -0.05 \rightarrow slope = Down \end{cases} \tag{1}$$

$$d = pitch_{end} - pitch_{avg} \begin{cases} \text{if } d > Pt \rightarrow end = Above \\ \text{if } (-Pt \leq d \leq Pt) \rightarrow end = At \\ \text{if } d < Pt \rightarrow end = Below \end{cases} \tag{2}$$

where Pt is the average $\pm$ 10, i.e. pitch average with a bit of tolerance for deviation.

3.2 Formalizing the Learning Problem

The main goal of the learning task is to differentiate silences in realtime based on partial information of an interlocutor's behavior (prosody only) and predict

the best reciprocal behavior. For best performance the system needs to find the right tradeoff between shorter silences and the risk of overlapping speech. To formulate this as a Reinforcement Learning problem we need to define states and actions for our scenario.

Using single-step Q-Learning the feature combination in the prosody preceding the current silence becomes the *state* and the length of the Silence Tolerance Window (STW) becomes the action to be learned. For efficiency we have split the continuous action space into discrete logarithmic values (see Table 1), starting with 10 msecs and doubling the value up to 1.28 seconds.

Table 1. Discrete actions representing STW size in msecs

Actions:	10	20	40	80	160	320	640	1280

The reward system used to support this learning problem needs to combine rewards from both length of silence and the occurrence of overlapping speech. We have come up with a reward scheme that encapsulates both. Rewards for decision that do not lead to overlapping speech are based on the size of the STW; a 10 msecs STW scores -10 while a STW of 1280 msecs scores -1280. This represents that shorter STW's are preferred over longer ones[4]. Overlapping speech is currently the only indicator that the agent made a mistake and decisions causing an overlap are rewarded -3000 points. To stimulate exploration of newly discovered actions all new actions start with estimated return at 0 points.

For accurate measurement of overlaps and silences the interviewing agent is equipped with 2 Prosody Trackers, one monitoring the interlocutor and the other monitoring its own voice. To reduce the number of actual generated overlaps due to erroneous trials, and to increase the available data that can be learned from, the agent gets a reward of -2000 for each canceled decision; this is used in situations where the selected STW turns out to be a tad too short. Another method of speeding up the learning is confining the learning space to only a few viable actions in the beginning, discovering new actions only as needed. This is assuming that there exists a single optimal STW for each pattern with degrading returns in relation to distance from that point; we do not need to explore a window of 40 msecs if a window of 80 msecs is considered worse than one of 160 msecs for a specific state. We start with only 2 available actions, 640 msec and 1280 msec, spawning new actions only as needed; spawning a smaller window only if the smallest available is considered the best and spawning a larger one if the largest tried so far is considered the best.

4 Experimental Setup

To evaluate the adaptability of our system we have conducted an experiment where the system, embodied as the agent Askur, automatically converses with

[4] We would like to thank Yngvi Björnsson for this insight.

10 human volunteers over Skype. Each subject is interviewed once, from start to finish, before the system goes on to the next subject. To eliminate variations in STW due to lack of something to say we have chosen an interview scenario, in which case the agent always has something to say until it runs out of questions and the interview is over. The agent is thus configured to ask 30 predefined questions, using silence tolerance window to control its turntaking behavior during the interlocutors' turn. Askur begins the first interview with no knowledge, and gradually adapts to its interlocutors throughout the 10 interview sessions.

A convenience sample of 10 Icelandic volunteers took part in the experiment, none of who had interacted with the system before. All subjects spoke English to the agent, with varying amounts of Icelandic prosody patters, which differ from native English-speaking subjects, and with noticable inter- and intra-subject variability. Each interview took around 5 minutes and the total data gathered is just over 300 turns of interaction (average 30 turns per subject). The study was done in a partially controlled setup; all subjects interacted with the system through Skype using the same hardware (computer, microphone, etc.) but the location was only semi-private and background noise was present in all cases.

The agent used the rewards, states and actions as described above, with expiration of STW decisions set to 1 second from the end of window; exploration was fixed at 10%.

5 Results

Given a group of people with a similar cultural background, can our agent learn to dynamically predict proper turn-transition silences during the dialogue. Can it adapt dynamically to each interlocutor during the interaction, while still improving as it interacts with more people? It can.

After having interacted with 8-9 people for about 30 turns each, our agent Askur achives human-level performance in minimizing pauses between turns: over 60% of turns are under 500 msecs and around 25% of turns are under 300 msecs. Furthermore, it learns to fine-adjust its turntaking behavior to a brand new person in under 30 turns. It does this while talking – no offline training trials are conducted. In our experiment the system achieves these levels in spite of interacting with non-native speakers of English. However, as our subjects' prosody patterns turned out to be significantly correlated, the agent keeps improving as more subjects interact with it.

Analysis of Askur's policy shows that out of 9 categories of prosody patterns one particular pattern has a much shorter Silence Tolerance Window (STW) than other patterns; only 38 msecs. This is for silences preceded with a final fall falling below average pitch (Down_Below) and is considerably shorter than the average length of Within-turn silence. Learned STW for all patterns can be seen in Table 2.

The data shows that Askur adapts relatively quickly in the very first 3 interviews, after which 50% of before-turn silences are shorter than 500 msecs (see Figure 2), compared to 70% in the human-human comparison data. 20% of silences are shorter than 300 msecs, compared to 30% within 200 msecs for the

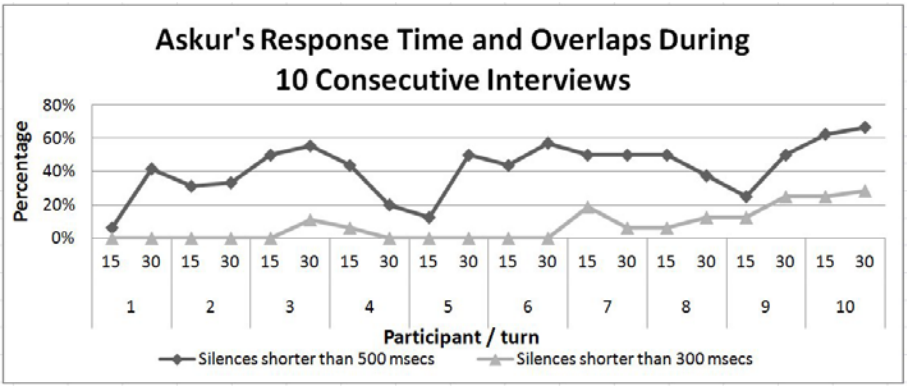

Fig. 2. Proportion of silences with human speed characteristics. The system interacts with each person for 30 turns, then switches to the next person for another 30, and so on for 10 sessions with 10 different people. For the first 3 interviews performance improves steadily, peaking at 60% of turn-transition silences under 500 msecs. Switching to a new person typcially does not impact the performance of the system. However, two obvious dips can be seen, for participant 4 and participant 8.

best-case human-human dialogue. Due to processing time in perception modules (73 msecs avg.) and "motor delay" in generating speech (206 msecs avg.) the agent never takes turn in shorter than 200 msecs (contrasted with 100 msecs for best-case human simple choice reaction time [22]). This performance in regards to response time is well acceptable and closely on par with human speed [1].

As can be seen in Figure 2 typically switching between people does not impair prior learning except for participants 4 and 8. We hypothesized that this might be due to interlocutor diversity, because of the unorthodox learning method of learning online while interacting with each subject sequentially. To investigate this hypothesis we analyzed the occurrences of the pattern Down_Below before silences in the speech of our 10 volunteers. The analysis shows that the occurrences vary between our volunteers from 2,7% to 41,03% of total occurrences at end of speech, and from 2,5% to 19,23% just before within-turn silences

Table 2. Learned Silence Tolerance Window (STW) based on prosody pattern

Prosody category	STW
Down_Below	38 msecs
Straight_At	160 msecs
Up_Below	320 msecs
Down_At	427 msecs
Straight_Below	440 msecs
Up_Above	480 msecs
Up_At	480 msecs
Straight_Above	640 msecs
Down_Above	640 msecs

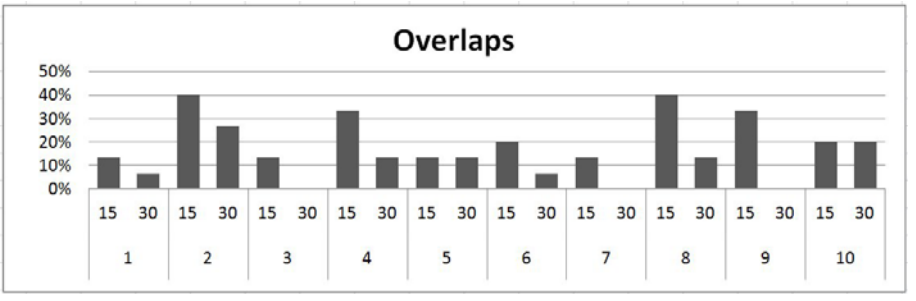

Fig. 3. Overlapped speech occurs on average in 26,3% of turns. Interestingly, overlaps periodically occur in 40% of turns without it permanently effecting performance and overlaps always decrease within each interview.

(see Table 3). This shows a considerable variation between subjects – even though subjects are all of same cultural background and speak the same (non-native) language in the sessions. Temporary lapses in performance are therefore to be expected. Yet the overall performance of the system keeps improving over the 10-person learning period. We can thus safely conclude that the subjects' diversity hypothesis is correct, and that the general trend shown by the data represent true learning on part of the system.

As can be expected in an open-mic conversation such as ours, overlaps were relatively frequent, with 26.3% of turns containing overlapping speech. This number, however, includes all simultaneous sound generated by each participant, regardless of whether it constituted actual speech, background noise or simply noise in the Skype channel. In the literature occurrence of overlapping speech has been found to vary considerably with type of conversation, observed to be as high as every 10th word in normal telephone conversations and telephone meetings [23] and 13% for normal conversation [24]. Interestingly, overlaps periodically occur in 40% of turns without it permanently effecting performance and overlaps always decrease within each interview (see Figure 3). 26.3% overlaps

Table 3. Usage of Down_Below per participant

Participant	turn-transitions	within-turn
1	7,69%	14,93%
2	14,81%	7,25%
3	34,78%	6,67%
4	6,25%	9,09%
5	2,70%	7,14%
6	27,27%	15,38%
7	41,03%	8,70%
8	18,75%	5,00%
9	11,11%	2,50%
10	25,00%	19,23%

when using prosody as the only information "channel" for determining turn transitions can thus be considered a success, especially given system's the continuous 10% exploration (we do not "train" the system and then turn learning off – learning is always on in our system).

5.1 Discussion

Research has shown that in casual conversation people adjust to the speaking style of their interlocutor, including the length of silences [4], producing a reasonably symmetric set of silences and overlaps for each participant. Our results show an asymmetry; Askur has in fact a noticably shorter duration for taking turn than the human subjects. This has a natural explanation, since in our dialogue Askur always has the role of an interviewer and the human are always in the role of the interviewee: The interviewer always knows what question to say next, whereas the human subject does not know what question comes next and has to think about what to answer. There is therefore typically a natural pause or hesitation while they think about what to answer to each question.

An important question that arises when learning on-the-fly against human subjects is whether the humans are actually participating in the learning performance of the system, essentially contributing to the learning by adapting their behavior to the system, consciously or unconsciously. If so this could conflate the results, reinforce "bad" behaviors of the system or otherwise bias the results. To answer this question we analyzed the data for cross-participant trends in modified behavior. While the use of filled pauses cannot be measured directly one way to detect them is to look at the duration of people's within-turn silences, which should be decreasing over time for each participant if the conflation hypothesis was correct. However, this was not the case: Average silence length stays constant throughout the interview for the participants (see Figure 4).

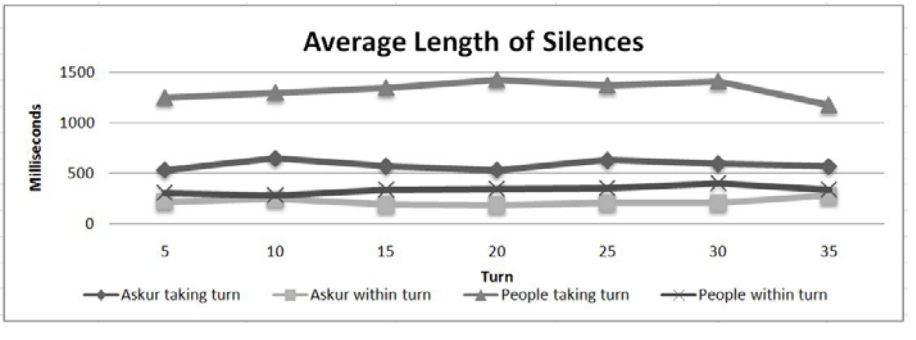

Fig. 4. Average silences when the agent interviews people. People's silences before taking turn are longer due to the fact that people have to think of what to say in response to the questions made by Askur. The length of silences within turn are constant throughout the interview verifying that people are not modifying their silences (by using filled pauses etc.) to accommodate the system.

In addressing this issue we also analyzed the use of final-fall preceding both within-turn and turn-transition silences. If people were learning to modify their use of final fall as a turn-giving signal we should see, towards the latter half of each interaction, a decrease in the occurrence of that pattern before within-turn silences and possibly increase before turn-transition silences. The data shows, however, that only 2 of the participants show this behavior while another 2 show the opposite behavior. The remaining 6 either decrease or increase the use at both within-turn and turn-transition silences (see Figure 5). No other common behaviors have been spotted that would suggest that the interlocutor is specifically aiding in the system's performance.

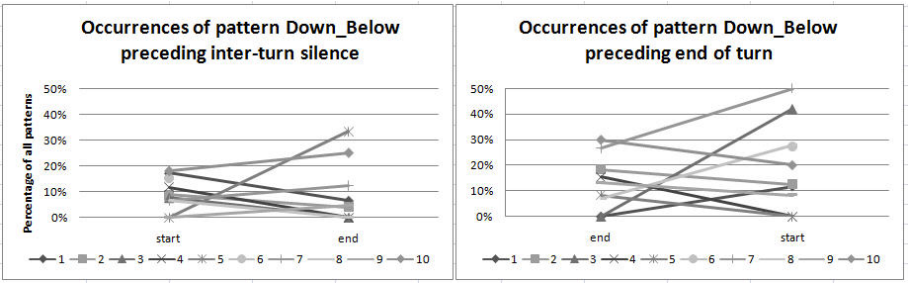

Fig. 5. Occurrences of Down_Below in people's speech has been measured for each interview. No trend in behavior is found that applies to majority of speakers.

6 Conclusions and Future Work

We have built a dialogue system that learns turntaking behaviors, minimizing silences and speech overlaps, using realtime prosody analysis. The system learns this on the fly, in full-duplex (open-mic) dynamic, natural interaction. The system can efficiently take turns with human-like timing in dialogues with people.

The system improves its performance by learning which prosodic information helps in determining appropriate turn transitions, combining features of pitch including final slope, average pitch and timed silences, as a evidence for predicting the desired turntaking behavior of interlocutors. As the system learns on-line it is able to adjust to the particulars of individual speaking styles.

We evaluated the system in realtime interaction with naive users. The system gets close to human-like speed when taking turns, with turn-transition silences as short as 235 msecs. 60% of turn-transition silences are shorter than 500 msecs after roughly 90 turns of learning, compared to 70% in human conversation.

In this evaluation all interlocutors were from the same cultural pool and thus had correlated intonation style, even though they were not speaking their native language. The learning system is embedded in a large expandable architecture, with significant potential for extensions beyond the interview scenario described here, including e.g. selecting dynamically between the goals of being polite

(no gaps, no overlaps) and "rude" (always trying to interrupt if it has something to say). As the system is highly modular it can be broadened to include multimodal perception such as head movements, gaze, and more.

In the near future we expect to expand the learning system to handle more diversity in interactions and variation between individuals. We also plan to expand the system to show dynamic selection of dialogue styles/goals such as politeness, agression and passivity. Semantic analysis of speech content will be integrated seamlessly with the turntaking capability, with a hope of going well beyond present-day dialogue engines in flexibility and human-likeness.

Acknowledgments. This work was supported in part by a research grant from RANNIS, Iceland. The authors wish to thank Yngvi Björnsson for his contributions to the development of the reinforcement mechanisms and Eric Nivel for the Prosodica prosody analyzer.

References

1. Wilson, M., Wilson, T.P.: An oscillator model of the timing of turn-taking. Psychonomic Bulletin Review 38(12), 957–968 (2005)
2. Ford, C., Thompson, S.A.: Interactional units in conversation: Syntactic, intonational, and pragmatic resources for the management of turns. In: Ochs, E., Schegloff, E., Thompson, S.A. (eds.) Interaction and Grammar, pp. 134–184. Cambridge University Press, Cambridge (1996)
3. Goodwin, C.: Conversational Organization: Interaction Between Speakers and Hearers. Academic Press, London (1981)
4. ten Bosch, L., Oostdijk, N., Boves, L.: On temporal aspects of turn taking in conversational dialogues. Speech Communication 47(1-2), 80–86 (2005)
5. Jefferson, G.: Preliminary notes on a possible metric which provides for a standard maximum silence of approximately one second in conversation. Conversation: an Interdisciplinary Perspective, Multilingual Matters, 166–196 (1989)
6. Thórisson, K.R.: Natural turn-taking needs no manual: Computational theory and model, from perception to action, pp. 173–207 (2002)
7. Thórisson, K.R., Benko, H., Arnold, A., Abramov, D., Maskey, S., Vaseekaran, A.: Constructionist design methodology for interactive intelligences. A.I. Magazine 25, 77–90 (2004)
8. Jonsdottir, G.R., Thorisson, K.R., Nivel, E.: Learning smooth, human-like turntaking in realtime dialogue. In: Prendinger, H., Lester, J.C., Ishizuka, M. (eds.) IVA 2008. LNCS (LNAI), vol. 5208, pp. 162–175. Springer, Heidelberg (2008)
9. Sacks, H., Schegloff, E.A., Jefferson, G.A.: A simplest systematics for the organization of turn-taking in conversation. Language 50, 696–735 (1974)
10. Walker, M.B.: Smooth transitions in conversational turntaking: Implications for theory, vol. 110, pp. 31–37 (1982)
11. Thórisson, K.R.: Dialogue control in social interface agents. In: INTERCHI Adjunct Proceedings, pp. 139–140 (1993)
12. Thórisson, K.R.: Communicative humanoids: A computational model of psychosocial dialogue skills, Ph.D. thesis, Massachusetts Institute of Technology (1996)
13. Gratch, J., Okhmatovskaia, A., Lamothe, F., Marsella, S., Morales, M., van der Werf, R.J., Morency, L.P.: Virtual rapport. In: IVA, Marina Del Rey, California, pp. 14–27 (2006)

14. Sato, R., Higashinaka, R., Tamoto, M., Nakano, M., Aikawa, K.: Learning decision trees to determine turn-taking by spoken dialogue systems. In: ICSLP 2002, pp. 861–864 (2002)
15. Schlangen, D.: From reaction to prediction: Experiments with computational models of turn-taking. In: Proceedings of Interspeech 2006, Panel on Prosody of Dialogue Acts and Turn-Taking, Pittsburgh, USA (September (2006)
16. Morency, L.-P., de Kok, I., Gratch, J.: Predicting listener backchannels: A probabilistic multimodal approach. In: Prendinger, H., Lester, J.C., Ishizuka, M. (eds.) IVA 2008. LNCS (LNAI), vol. 5208, pp. 176–190. Springer, Heidelberg (2008)
17. Bonaiuto, J., Thórisson, K.R.: Towards a neurocognitive model of realtime turntaking in face-to-face dialogue. In: Embodied Communication in Humans And Machines, pp. 451–483. Oxford University Press, Oxford (2008)
18. Thórisson, K.R., Jonsdottir, G.R.: A granular architecture for dynamic realtime dialogue. In: Intelligent Virtual Agents, IVA 2008, pp. 1–3 (2008)
19. Pierrehumbert, J., Hirschberg, J.: The meaning of intonational contours in the interpretation of discourse. In: Cohen, P.R., Morgan, J., Pollack, M. (eds.) Intentions in Communication, pp. 271–311. MIT Press, Cambridge (1990)
20. Thórisson, K.R.: Machine perception of multimodal natural dialogue. In: McKevitt, P., Nulláin, S.Ó., Mulvihill, C. (eds.) Language, Vision & Music, pp. 97–115. John Benjamins, Amsterdam (2002)
21. Nivel, E., Thórisson, K.R.: Prosodica: A realtime prosody tracker for dynamic dialogue. Technical report, Reykjavik University Department of Computer Science, Technical Report RUTR-CS08001 (2008)
22. Card, S.K., Moran, T.P., Newell, A.: The Model Human Processor: An Engineering Model of Human Performance, vol. II. John Wiley and Sons, New York (1986)
23. Andreas, E.S.: Observations on overlap: Findings and implications for automatic processing of multi-party conversation. In: Proceedings of Eurospeech 2001, pp. 1359–1362 (2001)
24. Markauskaite, L.: Towards an integrated analytical framework of information and communications technology literacy: from intended to implemented and achieved dimensions. Information Research 11 (2006), paper 252