

A Bidding Approach to Turn-Taking

Ethan O. Selfridge and Peter A. Heeman

Center for Spoken Language Understanding
Oregon Health & Science University
20000 NW Walker Rd., Beaverton, OR, 97006
selfridg@cslu.ogi.edu, heeman@cslu.ogi.edu

Abstract. We propose a new turn-taking framework for spoken dialogue systems in which conversants bid for the turn. This differs from most current turn-taking approaches, where the turn only changes after the holder has released it. Our new framework uses Reinforcement Learning to choose appropriate turn bids, which indirectly estimates the utterance importance. We evaluate this approach in an artificial task-oriented domain and we find that it outperforms conventional release-turn methods in a relatively realistic environment. Current performance is explained and future implications are discussed.

1 Introduction

Taking turns in speaking is a fundamental aspect of human conversation. It is essential for dialogue systems to engage in natural turn-taking behavior in order to facilitate efficient interaction with a human user. Turn taking in spoken dialogue systems (SDS) is often modeled quite simply; the turn only changes after the speaker has released it. This design is consistent with Sacks, Schegloff, and Jefferson [1], who state that the position for a possible turn release (transition relevance position; TRP) is highly predictable, and that conversants attempt to limit gaps or latencies between utterances while reducing utterance overlap by using this predictive quality. Though the assertion that TRPs are predictable has been criticized in the psycholinguistic field [2], the emphasis on turn release detection for transitions in dialogue systems has not diminished. Some current methods that predict the release-turn only use pause threshold. [3, 4]. Raux and Eskenazi [5] predict the release-turn using lexical cues, resulting in less latency between turns. Other work on reducing silence between turns uses Reinforcement Learning to tailor the onset of system utterances to a particular user [6]. Much of the work in turn-taking for embodied multi-agents systems [7, 4] utilize gaze, gesture, and other cues to signal the end of turn, in accordance with Duncan’s work on turn-taking [8].

Though there may be a predictive element to turn transitions, whether primarily syntactic as suggested by Sacks et al. or covering a much wider array of cues as proposed by Duncan, the systems mentioned above do not attempt to resolve turn-conflicts in any form besides releasing the turn.

We propose that a better approach to turn-taking, inspired by turn-conflict resolution, is a negotiative framework where the conversants bid for the turn. We call this *turn – bidding*. Our turn-bidding model is supported by empirical psycholinguistic research. In a study accurately predicting turn-assignment, Duncan and Niederehe [9]

indicated that the turn is won on the strength of one's turn cues. Yang and Heeman [10] found that the number of initiative conflicts grew as the amount of time the conversants had to speak was reduced. This temporal reduction is believed to increase the value or importance of the utterance due to the urgency to speak. Schlegoff [11] also proposed that longer utterance overlaps were due to competing "interest" of conversants in taking the turn.

We argue that utterance importance is a driving force behind turn-taking in general, not specifically for turn-conflict, and that these conflicts provide a window into the inner working of turn-assignment. In their study of discourse segments, Walker and Whittaker [12] suggest that misunderstandings between conversants often resulted in an interruption. This interruption, obviously important, would not necessarily be preceded by any release-turn cues. This suggests that, if the utterance is believed to be important enough, people do not wait for turn-release cues. It follows from this that people, wishing to speak, only limit their contributions due to insufficient conversational importance. We believe that people continually wish to speak, and that they compare the importance of their utterance with the turn-cues of the other conversant. If they believe their utterance to be important enough, they will try to take the turn. One could surmise that only crucial utterances will result in very strong turn cues ("Your house is on fire!"), the importance of the utterance outweighing the turn-cues of the speaker. Furthermore, conversants take the cost of overlap and turn-conflict resolution into account when deciding whether to take the turn or not.

Utterance onset during pauses is a strong indicator (among others) of utterance importance, conversants begin speaking quickly if they have something important to say. We design our prototypical turn-bidding model around differing onsets during pauses and we evaluate our design in an artificial environment. We compare it against two other trained systems: The first is a baseline single-utterance turn-taking model, each agent releasing the turn after speaking. The second system's framework is consistent with release-turn systems; two turn-actions (keep-turn and release-turn) are learned by the system agent, who can only attempt to take the turn after the user has released it.

2 Related Work

2.1 Keep-Or-Release Turn Approach

Many dialogue systems focus on the release-turn as the most important aspect of turn transitions, in which a listener will only take the turn after the speaker has released it.

Human-Human Studies Psycholinguistic research focuses on the importance of one speaker yielding the turn before another should take it. The conversation model proposed by Sacks, Schegloff, and Jefferson [1] has been extremely influential. Their model emphasizes the importance of conversational units called Turn Construction Units. A TCU can be a phrase, clause, sentence, or word, and it has a predictable end; a Transition Relevance Place (TRP). At a TRP, the speaker may select the next person to speak, a listener may self-select and begin speaking, or the speaker may continue speaking. Since the speaker controls the TRP, they control when another can take the

turn. Sacks et al. emphasize the importance of syntax to turn construction. Duncan [8] argues for the use of a wide array of turn-cues, offering non-syntactic cues such as gesture, falling pitch, and gaze direction change as other strong indicators of the turn release. He states that the turn transitions happen smoothly when the speaker gives off a number of turn-yielding cues that the listener attends to. When the listener recognizes that the speaker is yielding, they either back channel or take the turn. Duncan proposes that when the listener attempts to take the turn without the presence of turn-yielding cues they are disregarding this mechanism. He seems to create a dichotomy between regular turn-taking, which relies on these turn-yielding cues, and take-turn attempts, which are without said yielding cues. Both of these theories seem to agree on the function of release-turn for smooth transitions, which minimize silence and overlap.

Turn Actions in SDS There has been work done in spoken dialogue systems on defining a turn-taking mechanism that is consistent with the keep-or-release turn approach. Traum and Hinkelman [13] propose a series of turn-taking actions, namely take-turn, keep-turn and release-turn. The system can only use a take-turn action after it has determined that the user performed a release-turn action. In practice release-turns are usually identified by using a pause-threshold [3]. A variant of this three turn-action framework is used by Traum and Rickel [4], who add request-turn as a weaker version of take-turn. Since little resolution to turn-conflict is given we believe that the system will always release the turn when such a situation arises, if it arises at all. In a study using reinforcement learning in an artificial domain, English and Heeman [14] used only two turn-actions; keep and release. These two actions were learned by the two agents as the simulations progressed and neither agent could attempt to take the turn while the other still had it.

Release-Turn Systems Some systems do not rely on a pause-threshold and attempt to estimate the probability of a TRP occurrence based on combination of syntactic and prosodic cues. A successful TRP prediction would reduce the latency between utterance and allow for smoother dialogues. Kronlid [15] details the design of a turn-manager. This turn-manager, designed specifically to adhere the Sacks et al. model, has three primary components. One component handles the environment as a whole, another is concerned with overlap detection, and the last identifies TRPs. The agent is designed to stop speaking at overlaps and only begin speaking at TRPs. In his implementation, Kronlid suggests the use of the TRP predictor proposed by Hulstijn and Vreeswijk [16], in which the probability of the agent speaking is a ratio of the urgency of the agent to speak and the predicted distance to the TRP. Kronlid suggests that this urgency may be derived from “the status, self-confidence etc. of each agent”.

Raux and Eskanazi [5] continue on the TRP prediction vein, using decision theory to predict appropriate places to take the turn. They model the conversational floor as Jaffe and Feldstein did [17], as a 6-state finite-state machine. Like Kronlid, this model has a free state that operates as a legal transition position. Raux and Eskanazi compute the expected cost of taking the turn based on a cost matrix and the probability that the state is free. The cost is increased as the pause or latency between turns increases, and the probability of the free-state is akin to TRP prediction. They find that the use of this

cost function reduces the latency between utterances, the free-state probability being computed by logistic regression with lexical cues as the most informative feature.

Jonsdottir, Thorisson, and Nivel used Reinforcement Learning to reduce latency and minimize overlap between utterances [6]. Using overlap as negative reward and shorter predicted pause duration as positive reward, the learning agent is able to reduce the number of interruptions and average turn taking silence after approximately 40 artificial dialogues. Pause duration is used to decide whether to take the turn or not, so a shorter pause duration will lead to shorter latencies but more possibility of overlap. The authors state that since the agent is able to adjust relatively quickly, it can be used to accommodate users during the interaction and facilitate smooth human-like turn-transitions. While this approach is useful towards smoothing turn-transitions, it does not relate the transition to the dynamics of the dialogue which undoubtedly influence turn-taking.

2.2 Turn Resolution

The previous Section 2.1 assumed that turn-taking should be orderly. However, a number of researchers have drawn attention to turn-conflicts, in which the transition does not rely on release-turns and is distinctly disorderly.

Presence of Turn-Conflict Schlegoff [11] looked at the overlapping speech as a whole. He states that there is rarely more than two parties in overlaps even when there are many people in the conversation. He also draws a distinction between quickly resolved overlaps, and ones that persist quite longer, which can be characterized as turn-conflicts. The time that the conflict persists is a function of the interest that the conversants have in winning the turn, which is almost synonymous with utterance importance. Yang and Heeman [10] looked at turn-conflicts in the Multi-Threaded Dialogue (MTD) corpus. The MTD corpus is a collection of dialogues in which two humans play two games; one ongoing and task-oriented, and one called the interruption game, which was short, constrained by time, and played during the other game. They found that as the time constraint of completing the interruption game became shorter the number of turn-conflicts grew as well. More specifically, the percentage of conflicts grew from 9% at 40 seconds to 24% for 10 seconds. While analyzing discourse segmentation Walker and Whittaker [12] proposed a situation where interruptions, a type of turn-conflict, should occur. Their proposal expected interruptions when there was some discrepancy of mutual belief between conversants, and that this interruption facilitated conversational success. They seem to suggest that the previous turn-holder immediately released the turn, which is how most systems are currently designed to handle barge-in. However, there are other cases when both conversants compete for the turn and this must be resolved in a principled fashion.

Turn-Conflict Resolution Turn-conflict resolution is tantamount to turn-assignment. Schlegoff [11] suggests that turn-conflicts are resolved by competition. In a lengthy description of persistent competition, Schlegoff details the incrementally increasing strength of cues. He seems to propose that the conversant with the stronger relative

interest will exhibit the stronger turn cue. Yang and Heeman [10] also found that volume, specifically a relative increase in volume, was a strong indicator of winning the turn in utterance overlap situations. They found that they were able to correctly classify winners 79 % of the time using the higher relative volume. In line with the above research, Duncan and Niederehe [9] argued that the balance of turn-cues determined who won the turn. If a conversant gives more turn-taking cues than turn-releasing cues then they usually win the turn. Using this method, Duncan and Niederehe were able to correctly predict 18 out of 19 turn-assignments. Taken together, these studies support the dependence of turn-conflict resolution on turn-cue strength, which is dependent on one's interest in holding the turn.

3 Turn-Bidding Model

3.1 Psycholinguistic Framework

The previous Section 2.2 discussed the use of turn-cue strength to resolve turn-conflicts. In this paper, we contend that the driving force behind these turn-cues is an individual's interest in having the turn, and that it is natural to think that a person will be more interested in holding the turn if they believe their utterance to be important. Support for this importance-driven approach is given by the increase of turn conflicts during tighter time constraints found by Yang and Heeman [10], and conversant interest as detailed by Schlegoff [11]. In support of importance-driven turn-taking, research done by Walker and Whittaker [12] suggests that people will interrupt to remedy some understanding discrepancy, which is of obvious utility and importance to the conversation. We suggest that as the utterance importance grows so does the strength of their turn-cues, and that people are constantly using these cues to bid for the turn. Though conversants may be tracking turn-cues and utterance importance during speech, a natural turn-bidding situation is pauses. At pauses, the onset time of the utterance is an important component of bidding for the turn, the first one to speak winning the turn. It follows from this proposal that the amount of time that a conversant waits with starting their utterance (utterance onset) is dependent on utterance importance, and that conversants will use this onset to bid for the turn during pauses.

3.2 Computational Framework

We implement the psycholinguistic concept of turn-bidding outlined above in a dialogue system. The system, using utterance importance, bids for the turn after every utterance. Turn-bid actions are separate from "content" actions, and in this work we choose 5 bids (strongest to weakest): shorter, short, mid, long and longer. The winner of a tied bid is randomly decided.

Since we use an artificial environment the users also have turn-bids. We have two user classes: expert and novice. Expert users only use short bids and novice users only use long bids. Table 1 gives an example turn-bidding dialogue with an expert user. One can see that the conversant with the shortest bid wins the turn. An obvious challenge to this design is the measurement and assignment of importance. We use reinforcement

Table 1. Sample Turn-Bidding Dialogue with Expert User

Agent	Content Action	Sys Turn-Bid	Usr Turn-Bid
Sys:	query side		
		longer	short
Usr:	inform fries		
		mid	short
Usr:	inform burger veggie		
		mid	short
Usr:	inform drink cola		
		shorter	short
Sys:	Good Bye		

learning to overcome this. Reinforcement learning attempts to maximize the cumulative reward, and assigns a value to each action based on its contribution towards that maximization [18], and has been used effectively to develop dialogue systems (e.g. [19, 20]). In a dialogue setting, it is intuitive that an important utterance will lead to a higher reward, and so will have a higher bid value. This enables the system to assign importance to utterances based solely on the system’s experience, not some possibly arbitrary designer decision. We expect that the system will learn to couple higher-valued utterances with higher bids.

4 Task Specification

We use the information state update approach [21] for our agent and dialogue design. In this, each agent has an internal state that characterizes what the agent knows to be true about itself and its environment. This information state is made up of a number of domain variables, as well as dialogue processing variables such as ‘lastMove’, ‘haveTurn’, and ‘lastSpeaker’. At each turn a series of rules are applied to update the variables based on new information. We frame the agents and dialogue environment as a Markov Decision Process which allows for the use of standard reinforcement learning algorithms. We define our reinforcement learning variables as a subset of the information state variables, giving us both an Information State and a Reinforcement Learning State [22].

Two agents, a user and the system, hold a food-ordering dialogue (Table 1). The user wishes to order a specific type of burger, drink, and side and the system must take the order. We have two agent classes: experts and novices. The expert knows what foods are available and so always ask for items that are legal slot fillers. Novice users, on the other hand, must inquire on item availability if the system does not offer it. Sometimes the first choice of the novice user is unavailable and so they then inquire on their second choice or third choices.

The second turn-taking framework is based on the keep-or-release approach, which allows the system agent to learn whether to keep or release the turn and is modeled after the turn-taking framework of English and Heeman [14]. The system agent may not perform any actions unless the user has released the turn. In this setup, expert agents always keep the turn (unless they have already told the system all their slot values) and

the novice users always release it after making a contribution. Though these are imperfect analogs to turn-bidding users, we believe that it is close enough for comparative purposes. The third turn-taking framework we compare is a single-utterance approach where the system agent and both the expert and novice users release the turn after each utterances.

Ten policies are learned for each turn-taking framework with each of the three user environments. The first environment only includes expert users, the second only includes novice users, and the third includes an equal number of expert and novice users. In this third environment the system does not know which user it is interacting with. All together there are nine conditions. The dependent variable is the average converged cost (negative reward); a lower cost being indicative of a better policy. The dialogue cost is determined by the number of content actions and it is this that reinforcement learning strives to minimize. The turn-actions do not contribute to the cost. Each policy is trained for 1000 epochs, each epoch consisting of 100 dialogues. After each epoch the policy is updated. The policy is tested after each of the first 10 epochs, and tested every 20 epochs thereafter.

5 Results

We compare all three turn-taking frameworks by average converged cost, and policy comparison. Since the cost is only influenced by the number of content-actions, the higher the cost the worse the policy does. The average converged cost (Table 2) shows that across all three conditions turn-bidding performs as well or better than the keep-or-release approach. Turn-bidding outperforms both competing frameworks in the “Both” environment. This result is particularly important because the Both environment is the most realistic. Since dialogue systems have to interact with a variety of users they should be able to handle differences among them. Turn-bidding is able to adjust to interaction differences smoothly whereas keep-or-release is not. This performance was surprising since, due to simplicity of task and domain, we had expected the turn-bidding to meet but not exceed its competitor’s performance. These differences are discussed in greater detail below. Both Turn-Bidding and Keep-Or-Release did not perform quite

Table 2. Average converged cost over turn frameworks and user environments

Model	Novice	Expert	Both
Bidding	9.0	4.0	6.5
Keep-Or-Release	9.0	4.0	7.5
Single-Utterance	8.7	6.0	7.4

as well as the Single-Utterance model in the Novice user environment. The Single-Utterance model cannot learn a turn-taking strategy and so the dialogue length depends on the user. If the user’s slot values are available then the dialogue will be short, however if the user’s slots are not available then the dialogue will be long. The other two models can learn a turn-strategy to overcome these differences, and have a constant

dialogue length. However, the lack of strategy hurts the Single-Utterance model when handling Expert Users.

We analyzed the policies and found that the turn-bidding system learned to adjust its turn-bids for the importance of the utterance. For instance, when expecting a user to answer, the utterance onset is “longer” or “mid” (depending on the user), which allows the user to grab the turn. However, when it is of high value for the system to inform the user it uses “shorter”, such as when the user has asked a question. So, the aforementioned coupling appears to have been learned correctly. The keep-or-release system learns to generally release the turn in the expert environment and oscillate between keeping and releasing it in the novice environment. The standard release-turn system forces a turn transition after every utterance, resulting in a dampened information flow since it is unable to take advantage of consecutive utterances from the same agent. This causes far lower performance, as shown in Table 2.

Turn-bidding performs better than Keep-Or-Release in the ‘Both’ user environment, and we analyze the policy to explain this result. In ‘Both’, it is crucial to effectively handle novice users by informing them of filler availability after an initial slot query, while allowing expert users to inform the system on their first preference. By giving the availability utterance a mid level bid, the system is able to meet the needs of both user types while the keep-or-release framework does not allow for this functionality (Table 3). The keep-or-release system chooses to release the turn after the initial query which, while allowing the expert to inform the system, forces the novice to query the system and extends the length of the dialogue (Table 4).

Table 3. Turn-Bidding for “Both” Domain

Agent	Content Action	Sys Turn-Bid	Usr Turn-Bid
Sys:	query side		
Sys:	inform side have fries salad	mid	long
Usr:	inform side salad	longer	long

Table 4. Keep-or-Release for “Both” Domain

Agent	Content Action	Sys Turn-Action	Usr Turn-Action
Sys:	query side		
Usr:	query have salad side	release-turn	
Sys:	inform yes		release-turn
Usr:	inform side salad	release-turn	

6 Conclusion

Importance driven turn-bidding may make psychological sense but will it result in better dialogue systems? Using TRP prediction, the system never takes the turn while it believes the user still holds it and always releases the turn (i.e. stops speaking) when the system detects a user utterance. While this may make for pleased users in short dialogues, it is unclear whether this model will yield beneficial results for longer dialogues. We believe that the primary indicator of user satisfaction is efficiency, and modeling conflict is necessary for better efficiency. The conservative nature of the TRP focused systems may be detrimental on this front since it may halt important utterances in deference to the aggressive user. On the other hand, a turn-bidding agent would not yield the turn to a speaking user unless its utterance was less important than the user's. Furthermore, a turn-bidding agent will be able to attempt to take the turn if it believes its utterance to be important, this being crucial for conversational success [12] This may be a more useful and simple framework than the inherently inflexible release-turn dialogue agents.

Since the turn-bidding system is able to couple high-valued utterances with short onsets, we see the success of reinforcement learning to operationalize importance. The utterance value learned by RL can be directly related to a global dialogue importance since it is focused on maximizing the dialogue reward as a whole. Since reward in this case is primarily characterized by dialogue length, one can easily see that objective functions with greater sophistication could lead to agents with even more intricate patterns of interaction. Our use of RL differs from that of Jonsdottir et al. [6] in that we are optimizing on the entire cost of the dialogue, not just the factors affecting the turn transition.

These results show that the use of reinforcement learning with the turn-bidding approach does not result in conversational deficiencies when compared to conventional release-turn systems. In the more realistic 'Both' environment, turn-bidding improves on the competing Keep-Or-Release framework due to its ability to handle multiple user classes. We are currently improving the turn-bidding framework, and hope to show that it continues to outperform conventional methods when used in complex multi-agent environments and when used with real users.

Acknowledgments The authors gratefully acknowledge funding from the National Science Foundation under grant IIS-0713698.

References

1. Sacks, H., Schegloff, E., Jefferson, G.: A simplest systematics for the organization of turn-taking for conversation. *Language* **50** (1974) 696–735
2. Connell, D., Kowal, S.: Turn-taking. In: *Communicating with One Another*. Springer-Science + Business (2008) 149–161
3. Sutton, S., Novick, D.G., Cole, R., Vermeulen, P., de Villiers, J., Schalkwyk, J., Fenty, M.: *Building 10,000 spoken-dialogue systems*, Philadelphia (1996)
4. Traum, D., Rickel, J.: Embodied agents for multi-party dialogue in immersive virtual worlds. In: *Proceedings of Autonomous Agents and Multi-Agent Systems*. (2002) 766 – 773

5. Raux, A., Eskenazi, M.: A finite-state turn-taking model for spoken dialog systems. In: Proceedings of HCL/NAACL 2009, Boulder, Co (2009) 629–637
6. Jonsdottir, G.R., Thorisson, K.R., Nivel, E.: Learning smooth, human-like turntaking in realtime dialogue. In: IVA '08: Proceedings of the 8th international conference on Intelligent Virtual Agents, Berlin, Heidelberg, Springer-Verlag (2008) 162–175
7. Cassel, J., Bickmore, T., Vilhjalmsson, H., Yan, H.: More than just a pretty face: Affordances of embodiment. In: Proceedings of the 5th international conference on Intelligent user interfaces. (2000) 52 – 59
8. Duncan, S.: Some signals and rules for taking speaking turns in conversations. *Journal of Personality and Social Psychology* **23** (1972) 283–292
9. Duncan, S., Niederehe, G.: On signalling that it's your turn to speak. *Journal of Experimental Social Psychology* **10** (1974) 234–247
10. Yang, F., Heeman, P.A.: Initiative conflicts in task-oriented dialogue". *Computer Speech Language* **24** (2010) 175 – 189
11. Schegloff, E.: Overlapping talk and the organization of turn-taking for conversation. *Language in Society* (**29**)
12. Walker, M., Whittaker, S.: Mixed initiative in dialogue: an investigation into discourse segmentation. In: Proceedings of the 28th Annual Meeting of the Association for Computational Linguistics. (1990) 70–76
13. Traum, D., Hinkelman, E.: Conversation acts in task-oriented spoken dialogue. *Computational Intelligence* **8** (1992)
14. English, M., Heeman, P.: Learning mixed initiative dialog strategies by using reinforcement learning on both conversants. In: Proceedings of HLT/EMNLP. (2005) 10111018
15. Kronlid, F.: Turn taking for artificial conversational agents. In: Cooperative Information Agents. Springer-Verlag Berlin Heidelberg (2006) 81–95
16. Hulstijn, J., Vreeswijk, G.: Turntaking: a case for agent-based programming. Technical report, Institute of Information and Computing Sciences, Utrecht University (2003)
17. Jaffe, J., Feldstien, S.: Rhythms of Dialogue. Academic Press (1970)
18. Sutton, R., Barto, A.: Reinforcement Learning. MIT Press (1998)
19. Walker, M.: An application of reinforcement learning to dialogue strategy selection in a spoken dialogue system for email. *Journal of Artificial Intelligence Research* **12** (2000) 387–416
20. Levin, E., Pieraccini, R., Eckert, W.: A stochastic model of human-machine interaction for learning dialog strategies. *IEEE Transactions on Speech and Audio Processing* **8** (2000) 11 – 23
21. Larsson, S., Traum, D.: Information state and dialogue management in the trindi dialogue move engine toolkit. *Natural Language Engineering* **6** (2000) 323–340
22. Heeman, P.: Combining reinforcement learning with information-state update rules. In: Proceedings of the Annual Conference of the North American Association for Computational Linguistics, Rochester, NY (2007)