Using Transition Duration to Improve Turn-taking in Conversational Agents

Charles Threlkeld and **Muhammad Umair** and **JP de Ruiter**
Tufts University
{charles.threlkeld, muhammad.umair, jp.deruiter}@tufts.edu

Abstract

Smooth turn-taking is an important aspect of natural conversation that allows interlocutors to maintain adequate mutual comprehensibility. In human communication, the timing between utterances is normatively constrained, and deviations convey socially relevant paralinguistic information. However, for spoken dialogue systems, smooth turn-taking continues to be a challenge. This motivates the need for spoken dialogue systems to employ a robust model of turn-taking to ensure that messages are exchanged smoothly and without transmitting unintended paralinguistic information. In this paper, we examine dialogue data from natural human interaction to develop an evidence-based model for turn-timing in spoken dialogue systems. First, we use timing between turns to develop two models of turn-taking: a speaker-agnostic model and a speaker-sensitive model. From the latter model, we derive the propensity of listeners to take the next turn given TRP duration. Finally, we outline how this measure may be incorporated into a spoken dialogue system to improve the naturalness of conversation.

1 Introduction

Turn-taking is an important component of many spoken dialogue systems and involves (a) detecting or predicting the end of a turn and (b) accurate timing of the initiation of speech production (Michael, 2020; Kennington et al., 2020). Smooth turn-taking continues to be a challenge for spoken dialogue systems that aim to engage in natural conversation (Hara et al., 2019). Traditionally, most spoken dialogue systems process Inter-Pausal Units (IPUs), which are speech units surrounded by arbitrary fixed length silence thresholds (Skantze, 2021). These pauses of arbitrary duration cause a stilted, unnatural conversation style.

Other systems use incremental approaches, processing smaller units of speech at a time. For example, the incremental dialogue system proposed by Skantze and Schlangen (2009) operates on Incremental Units (IUs) that are processed by Incremental Modules (IMs). These modules may include action, turn, or dialogue management—each of which influences turn-planning and end of turn detection. Although such systems initiate the production of speech when a silence is detected, they do so based on pitch or semantic completeness (Skantze and Hjalmarsson, 2010). Machine learning models of turn-taking operate on previously detected multimodal cues (Bohus and Horvitz, 2010; Skantze, 2021).

The turn-taking techniques used in traditional spoken dialogue systems, such as predicting turn-ends in a time window (Lala et al., 2019), are not fully grounded in current theory of human turn-taking. In natural conversation, people tend to minimize gaps and overlaps while also following the one "one speaker at a time" rule (Sacks et al., 1974). This means that when a turn ends, another speaker may start speaking. Speakers also use the duration of silences to convey social information (de Ruiter, 2019). For example, long and short gaps may communicate hesitance or impatience. Additionally, interlocutors use turn-taking cues (e.g., lexico-syntactic, pragmatic, prosodic etc.) to predict the end of turns and plan responses (Levinson and Torreira, 2015; Liddicoat, 2004). In contrast, turn-taking techniques typically do not explicitly identify points where floor change may occur (Hara et al., 2019), normatively time the duration of silences, or predict turn-ends independent of the occurrence of specific events (e.g, silences) (Skantze and Hjalmarsson, 2010). For spoken dialogue systems, this leads to mistimed responses and a decrease in human engagement (Zhao et al., 2018).

In this paper, we propose an evidence-based model for *when* speech may be produced to facilitate smooth turn-taking, based on the turn-taking model proposed by Sacks et al. (1974). In this model, a speaker's turn consists of one or more

Proceedings of the SIGdial 2022 Conference, pages 193–203
Heriot-Watt University, Edinburgh, UK. 07-09, September, 2022

Turn Construction Unit (TCU), which encompasses sentential, clausal, phrasal, and lexical constructions. Between each TCU are Transition Relevance Places (TRPs), where the current turn may be completed and a floor change may occur (Selting, 2000). We further divide TRPs into continuations (TRPs where the current speaker continues) and switches (TRPs where a speaker-transition occurs). Additionally, in our operationalization, each TRP has a duration—the time between when the previous TCU is complete but before the next TCU begins. We use TCU-level data from transcriptions of the Switchboard corpus (Godfrey and Holliman, 1993) to develop two models of turn-taking based on the duration of TRPs: a speaker-agnostic model and a speaker-sensitive model. Next, we develop an evidence-based function for the propensity of floor-transfer as a function of time after the end of a TCU. Finally, we outline a proposal for implementing this propensity function into the continuous dialogue system architecture formalized by Skantze and Schlangen (2009).

2 Motivation and Related Work

2.1 Conceptual Models of Turn-Taking

Two models of turn-taking have been proposed in the turn-taking literature: Duncan's signal-based model (Duncan, 1972) and Sacks, Schegloff, and Jefferson's "simplest systematics" model (hereafter the Sacks et al. model) (Sacks et al., 1974).

Duncan's model of turn-taking proposes that speakers produce turn-keeping and turn-yielding signals that are picked up by listeners, thereby ensuring smooth floor transfer. Turn-yielding signals include, among others, changing intonation, specific syllable stress patterns, and gesture ending or relaxation.

Previous work has shown that intonational phrases at unit boundaries are signals used in end-of-turn detection (Bögels and Torreira, 2015a). Gravano and Hirschberg (2011) found that the greater the number of turn-end cues present in a phrase, the greater the likelihood of a floor transfer occurring. Similarly, Ford and Thompson (1996) found that syntactic, intonational, and pragmatic completeness are all required for smooth turn transition. One important takeaway from this model is that the *speaker* yields the turn, and is therefore the main decision maker for whether turn transition occurs. Speakers can therefore control whether the listener takes over the turn or not.

In contrast, Sacks et al. (1974) propose a model of turn-taking in which listeners can (but do not have to) take the floor at so-called Transition-Relevance-Places (TRPs). According to Sacks et al., listeners can predict (or "project") ahead of time when the TRP will occur. Once a TRP has been reached, the rules specified by Sacks et al. are that a) the current speaker may select the next speaker, b) if that does not happen, a next speaker may self-select, and c) if no speaker self-selects, the current speaker can continue.

2.2 Conceptual Implications

Interestingly, the differences between these models of turn-taking make different predictions as to the duration of the TRP as a function of whether the same or a different speaker takes the floor. In the Duncan model, the speaker controls the floor transfer using signals, allowing them to keep the rhythm of the conversation steady. In contrast, in the Sacks et al. model, when a speaker arrives at at TRP and has not selected the next speaker, the speaker can only continue their turn after having established that the listener did not self-select. This predicts that if the Sacks et al. model is correct, we will see shorter TRP durations when there is a speaker change than when there is not.

Therefore, the Sacks el al. model predicts that there are two separate distributions of TRP duration: one for TRPs at speaker switches and the other for TRPs at continuations. Since, according to the rules, a listener has the first option for uptake during a TRP, we expect that the TRP duration for speaker switch is faster than for speaker continuation. Of course, the probability distributions are likely to overlap. A speaker may sometimes continue with a small pause or there may be a long pause before a speaker switch. We interpret this model to mean that the speaker switch distribution will be generally faster than the speaker continuation distribution.

There is a third possibility: that speaker continuation is faster than speaker switch. While neither model predicts this, it could happen, for instance, if we assume the Duncan model is correct, and listeners do not detect the turn-yielding cues, or detect them too late. This implies if we find speaker continuations to be faster than speaker switches, it will be evidence for Duncan's model, and against Sacks et al.'s model.

2.3 Application in Spoken Dialogue Systems

Detecting the end of turns and timing speech production is vital for a spoken dialogue system to engage in smooth turn-taking. Accordingly, there are a number of approaches for automated end of turn detection in existing literature (e.g., Masumura et al. (2018, 2019)). A number of approaches have also been proposed for quick turn-transitions in spoken dialogue systems. For example, Gervits et al. (2020) found their incremental model was ready to reply to an utterance 635 ms ($\pm 197ms$) before the end of a turn. Since the mean gap between turns is generally between 0 ms and 200 ms (Heldner and Edlund, 2010; Stivers et al., 2009), this leaves an agent with significant temporal space within which to decide when to start turn production. Our goal, in contrast, is to address the characteristics of *naturalness* in smooth turn-taking timing (Edlund et al., 2008). Therefore, we assume an existing model for end of turn detection and propose a extension module of natural turn taking timing in spoken dialogue systems.

3 Empirical Models

In this section, we fit two Bayesian models of TRP duration: one that assumes a single distribution of all the TRPs in a dialogue, and one that assumes that the distribution is different for speaker switches and speaker continuations.

In what follows, we will first describe the empirical data that forms the basis for our models. Next, we provide a detailed description of the two probabilistic models informed by our conceptual models. We then describe the implications of our findings for turn-taking and speaker selection. Finally, we propose an evidence-based turn-taking propensity function for natural speech production decisions after the end of a TCU.

3.1 Data

We are interested in the duration of TRPs i.e., the timing between TCUs, in natural dialogue. Therefore, our data must consist of dialogue with TCU-level segmentation and highly accurate timing (down to the millisecond). We gathered this information from two different transcriptions of the Switchboard corpus—a corpus of dyadic telephone conversations (Godfrey and Holliman, 1993). The Mississippi State University transcriptions (MSU)[1] provide word-by-word timing, which has been hand-corrected to reduce word error rates to below 1%. The Switchboard Dialogue Act Corpus (SwDA)[2] segmented the Switchboard corpus into TCUs in order to annotate dialogue acts.

The Switchboard corpus is appropriate for this task because participants do not have access to many of the cues of face-to-face interaction (Duncan, 1972). This simplifies the work to only account for spoken language. Although using a corpus of telephone conversations may limit the applicability of our work in face-to-face interaction, it is appropriate for systems where this information is not available (Bosch et al., 2004).

Here, we outline the preprocessing steps applied to the data for the work presented in this paper. First, we merge MSU and SwDA transcripts of the same conversation to create a subset of the Switchboard corpus with transcriptions segmented at the TCU-level and annotated with accurate timing information. From this subset, we selected only conversations where the exact word-level matches were at least 90% of the words in the conversation and the total uncaught word error rate was below 2%. This allowed us to maintain data quality and yielded 75 conversations with acceptable timing information.

Next, we filter our timing data based on the following reasons. Our data consists of the duration of TRPs between TCUs. This duration may be positive or negative for speaker-switch TRPs (i.e., pauses and overlaps). To make a reasonable comparison between values of two different domains, we fit a truncated distribution to the data. For an overlap, we know that a speaker may not overlap with oneself and there is an obvious 'barge-in' when the overlap occurs. Therefore, we set a TRP floor above 0 ms. Further, previous research shows that pauses of one second or longer may be considered trouble sources in conversation (Jefferson, 1983; Roberts et al., 2006). Trouble source detection is out of the scope of this work. Therefore, we removed all TRPs with a duration greater than 1000 ms, which has the additional benefit of removing outliers from the data that might have skewed our models. Finally, we have two datasets: one for speaker switch TRPs and one for speaker continuation TRPs.

We recognize that this subset of data excludes overlaps, which are common in natural dialogue. However, this paper reasons about turn-taking

[1]The MSU corpus is hosted on OpenSLR.

[2]The SwDA corpus is available through Stanford.

through the duration of a silence. Our models do not make claims regarding when speech reasoning may occur, and instead focus on resultant behaviors that are exhibited [3].

The models and analyses below are based on the 4563 TRPs in our filtered dataset. 2686 of these TRPs were followed by speaker switches and 1877 were followed by speaker continuations. All models described are Bayesian models using truncated normal distributions with lower bounds of 0 ms and upper bounds of 1000 ms, in order to conform to the assumptions outlined above. The models were fit using `pymc3` version 3.11.4, a probabilistic programming package for Bayesian modeling. All priors[4] were designed to be weakly informative based on previous research in the field(Stivers et al., 2009; de Ruiter et al., 2006b). Weakly informative priors are also considered best-practice when using Markov-chain Monte Carlo (MCMC) Bayesian updating (Lemoine, 2019).

3.2 Speaker-Agnostic Model

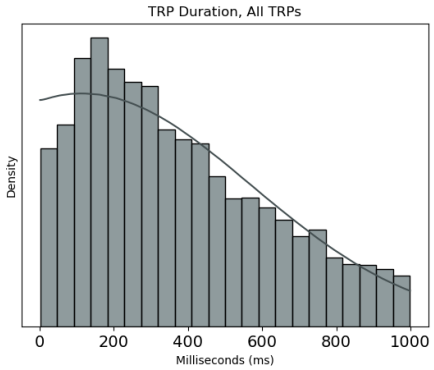

Figure 1: This figure shows a histogram with 50 ms bins of all TRPs with duration between 0 ms and 1000 ms. The best-fit truncated normal curve line is also shown.

The speaker-agnostic probabilistic model assumes that TRPs have a single underlying distribution. The assumption is that all TRPs are a function of the rhythm of the dialogue, controlled by the speaker, and pause durations are not influenced by who was speaking before the pause. Under this model, when there is a pause in the conversation, each participant has the same chance of deciding to continue.

As shown in Figure 1, the estimated mode TRP duration under these assumptions is in the 150–200 ms bin and the mean is 374 ms. The mean TRP duration, according to the posterior predictive model, is 375 ms. Since we are fitting a truncated normal distribution, the mean will be larger than the mode because the distribution is right-skewed. We see this in both the data and the model. The standard deviation of both the data and the posterior predictive model is 250 ms—indicating that the model has a good fit when assessed using the first two statistical moments. It is interesting to note that the mode of our empirical data is in the range 150–200 ms. This is the same mode range that previous work has determined for floor transfer offset, based only on the speaker switch condition (Levinson and Torreira, 2015; Heldner and Edlund, 2010; Stivers et al., 2009). Note that this previous research has focused on floor transfer for entire turns only, which is easier to operationalize since it does not require segmenting turns into TCUs.

	$\mu_{\text{hdi 3\%}}$	μ_{mean}	$\mu_{\text{hdi 97\%}}$	σ_{mean}
μ	56	110	167	30
σ	421	458	495	20

Table 1: This table describes the parameters of the best-fit truncated normal curve for all TRPs greater than 0 ms and no more than 1000 ms. The high-density intervals are given for a 94% high density interval i.e., the models predict only 3% probability that the true values lie above or below these intervals. The σ_{mean} terms are another measure of confidence, though not sensitive to skew.

3.3 Speaker-Sensitive Model

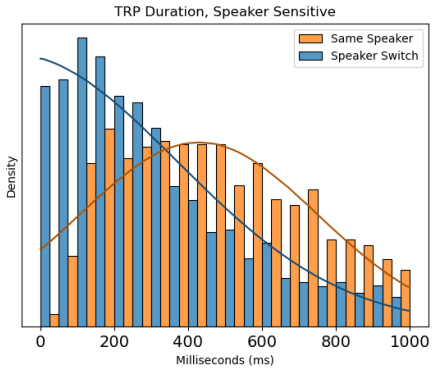

Figure 2: This figure shows a histogram of the empirical TRP data from 0 ms to 1000 ms broken down into 50 ms bins and in two conditions: speaker switch and no speaker switch. The best-fit truncated normal distribution lines for each condition are also shown.

We use the speaker-sensitive probabilistic model

[3] Alternatives to our models are in Appendix A.3.

[4] Prior distributions useds can be found in Appendix A.1.

to test the prediction from the Sacks et al. model. If this model is correct, we expect to see a different TRP distribution for the continuation and switch conditions. This is because the rules specified in their model lead to shorter TRPs when there is a speaker switch than when the same speaker continues, because the speaker first has to wait to see if a speaker self-selects before continuing their turn.

	$\mu_{\text{hdi 3\%}}$	μ_{model}	$\mu_{\text{hdi 97\%}}$	σ_{model}
μ_{switch}	-164	-85	-10	41
σ_{switch}	417	451	488	19
$\mu_{\text{continuation}}$	407	428	447	11
$\sigma_{\text{continuation}}$	300	323	347	13

Table 2: These statistics describe best-fit truncated normal curves for the independently fit curves. $switch$ variables are for cases where a speaker switch occurs at a TRP. The $continuation$ condition has the same speaker before and after the TRP. The high-density intervals are given for a 94% interval i.e., the models predict only 3% probability that the true values lie above or 3% below these intervals. Similarly, the σ_{model} terms are a measure of confidence that is not sensitive to skew.

In our speaker-sensitive model, we fit two distributions: one for speaker switch and one for a speaker continuation. The data contains 2686 TRPs where the speaker switches and 1877 TRPs where the speaker continues for all pauses in conversation from 0 ms to 1000 ms. We expect shorter pauses to be followed by a different speaker while longer pauses are followed by the same previous speaker. A fast speaker switch entails understanding and uptake by the interlocutor. A pause and continuation, in contrast, gives space for the listener to take the floor before the current speaker continues their own turn.

We found distributions that are substantially different for the speaker switch and continuation conditions. The Kolmogorov-Smirnov statistic for the two categories is 0.289 ($p < 0.01$). In the speaker switch condition, the mean of the best-fit posterior predictive is 315 ms, a bit slower than the data mean of 311 ms. The mode of the data shows that the floor transfer pause duration is 100–150 ms when binned into 50 ms segments, which aligns with previous work. This means that there is a preference toward fast responses. As a reminder, we filtered out any overlapping speech since it is outside the scope of this paper, even though we want to note that work on floor transfer offset (e.g., de Ruiter et al. (2006a); Heldner and Edlund (2010); Riest et al. (2015)) shows that overlap is a common phenomenon.

For speaker continuation, the data mean is 462 ms and the posterior predictive model mean is 459 ms. The mode of the data is 150–200 ms, although the values are very close for many other bins from about 100 ms to 600 ms, as shown in Figure 2. These data align closely to the predictions made by Sacks et al.'s model—speaker switch happens quite quickly, and speaker continuation somewhat later.

3.4 Model Selection

In Sections 3.2 and 3.3, we defined and fitted two Bayesian models to test differential predictions of the conceptual models presented in Section 2.1. We showed that each empirical model has a good quantitative and qualitative fit with the empirical data. We now want to know whether the model that incorporates speaker information is better even if we take into account that it has an extra parameter.

We will use linear mixed effects regression models, which are a common tool for differentiating trends based on groups within a population. We created two models, one with a speaker switch included, and one without. To account for different aspects of individual conversations, both models included the conversation identifier as a random factor. We used the `rstanarm` package in R (Goodrich et al., 2020) because it allowed us to use *bridge sampling* (Gronau et al., 2020) to compute Bayes Factors for the purpose of model comparison.

Our analysis shows that the data is 1.43×10^{80} times more likely under the speaker-sensitive model than under the speaker-agnostic model, even when correcting for the higher model complexity of the speaker-sensitive model. This constitutes decisive evidence (Wetzels et al., 2011) for next-speaker being influenced by TRP duration, supporting Sacks et al.'s model of turn-taking. Our results have two implications: (1) when responding, gaps should be minimized so that the speaker does not take the silence as an invitation to continue their own turn, and (2) after speaking, a response should come within the first few hundred milliseconds. Any longer, and the speaker may want to continue their own turn to maintain progressivity (Stivers and Robinson, 2006).

3.5 Turn-taking Propensity Function

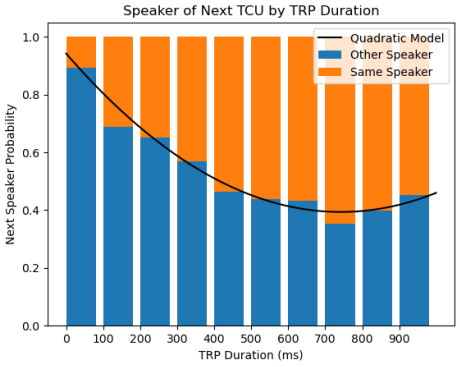

Figure 3: This figure shows the proportion of speaker-switch vs. speaker-continue events for each 100 ms TRP bin between 0 ms and 1000 ms, along with the best-fit quadratic line.

We have described two conceptual models of how turn-taking works, built probabilistic models based on these conceptual models, and established that the speaker-sensitive model inspired by Sacks et al. (1974) fits the data much better. We will now explore how we can use this knowledge to improve *when* conversational agents initiate their turn. To answer this question, we have one missing piece: we need to determine the propensity for speaker switch as a function of TRP duration. Note that the speaker-sensitive model we have formulated can be seen as two separate models: one for speaker switch and one for continuation. As mentioned before, the speaker switch condition was generally more frequent: 2686 TRPs with speaker switch and 1877 with speaker continuation in our dataset. In this section we will explore the relative proportion of speaker switch and continuation as a function of transition time.

Figure 3 shows the proportion of speaker switch and speaker continuations in our data. It shows that, as a silence grows longer, the relative propensity for a speaker to continue initially increases, while the relative propensity for a speaker switch decreases. However, as the silence continues, the share of floor holding decreases. We fit a basic quadratic curve to the floor transfer trend shown in Figure 3. The function below gives the maximum likelihood estimate for probability of speaker-switch as a best-fit quadratic function of the number of milliseconds of silence (t) since the previous turn ended.

$$P_{\text{switch}} = (9.70\times10^{-7})t^2 - (1.48\times10^{-3})t + 0.933$$

It is important to note that our analysis only looks at the first second of silence after a TCU. We did perform a cursory exploration of longer pauses, to check if there were obvious trends. We found that for silences between 1000 ms and 2500 ms, 56% of TCUs were floor-hold (speaker continuation). We caution against over-interpretation of these numbers, as there were 4563 TRPs between 0 ms and 1000 ms, but only 672 between 1000 ms and 2500 ms. Gaps longer than a second in conversations are rare in conversation (Jefferson, 1983), and may have a variety of causes.

A spoken dialogue system can use this formula and our two empirical models above in two ways: (1) timing its own responses and (2) setting response seeking limits. Current techniques allow for extremely fast response rates in spoken dialogue systems. An agent implementing our models can choose times that are acceptable to human dialogue speed. These times do not need to rely on heuristics like the mean FTO or barge-in mechanics, but can keep conversation at a fluid and natural pace on an utterance-by-utterance basis. Our model suggests that an agent should respond to a turn within 394 ms—the point at which each speaker has equal propensity to speak—ideally around 150–200 ms after a turn end, where the probability of a speaker change is still close to maximal.

A spoken dialogue system can also incorporate the propensity function during language generation to make sure that its turn-internal pauses are not too long or too short. If the system knows it wants to continue the turn in a subsequent TCU, it should flow fluidly, rather than give space for the interlocutor to respond.

Finally, if the planned turn is over, an agent could set a maximum listening time, after which it prompts a response or clarifies its previous statement. Our findings show that the agent should aim to do this around 762 ms, the minimum point of our speaker-switch function. Pauses of longer than a second are signs of trouble in a conversation, so continuing a turn is preferable than waiting indefinitely for a response (Jefferson, 1983; Roberts et al., 2006). Adding this functionality to a spoken dialogue system will provide an agent with the ability to ensure that the conversation progresses, and even prompt an interlocutor if they are unresponsive.

The function presented here is meant to be a baseline for turn-taking mechanisms. There are clear paths for extending it, like sensitivity to di-

alogue acts, ellipses, or prosody, but the overall effect should be similar in aggregate since the data here is presented in aggregate.

4 Continuous Module Proposal

In this section, we outline a proposal for operationalising our turn-taking propensity function for timing turn-taking. First, we outline relevant components of the incremental dialogue processing architecture proposed by Schlangen and Skantze (2011). Next, we define the *minimal* module implementing the proposed timing method, as well as possible extensions to incorporate existing turn-taking methods (e.g, Bögels and Torreira (2015b)).

We use the Schlangen and Skantze (2011) architecture for its continuous and incremental properties, which may be useful for comparing different timing methods. However, our proposed method is based simply on timing and does not have a strong dependence on any specific architecture.

4.1 Spoken Dialogue System Architecture

The conceptual model of incremental processing described in Schlangen and Skantze (2011) has two basic components. Incremental Units (IUs) are the basic units of processing and contain payloads (e.g., audio streams, words etc.) that can be processed by Incremental Modules (IMs). Each IM has a Left Buffer (LB) to store incoming IUs and a right buffer to store outgoing IUs. An IM also has a processor that consumes LB IUs and produces RB IUs. IMs communicate with each other by adding IUs to their RB, which is immediately available for LB consumption of connected IMs. Note that the rate of RB IU production does not need to match the rate of LB IU consumption.

Additionally, IUs may be connected to one another using relations, which effectively track the flow of information throughout the system. While there can be many different types of relations, we introduce two—spatial and grounded-in. The spatial relation connects IUs produced by a single IM. For example, for an IM generating turns from words, spatial links may be used to connect words that form the same turn. Second, the grounded-in relation can be used by IMs to connect RB IUs to their corresponding LB IUs. For example, this may allow a word recognized by an Automatic Speech Recognition (ASR) IM to be connected to the corresponding audio signal.

Finally, both IUs and IMs contain specified properties and operators. Each IU contains basic metadata type information that may be used for decision-making in individual IMs. This includes information for an IU to indicate its relations with other IUs, the confidence of the IM in the IU data, whether the IU result is final, and whether the IU has been processed by a specific IM. Similarly, IMs must implement certain methods including a purge method to reset the module's internal state, a new IU update method to update the module state based on incoming information, and a commit method to finalize the IUs in its RB.

4.2 Module Incremental Units

Our proposed module aims to provide more granular turn-taking timing information to the spoken dialogue system compared to existing approaches, which plan and execute entire turns (Jokinen et al., 2013). Therefore, it produces RB IUs whose payloads are waiting times after which the dialogue system should take the next turn. Additionally, the IU includes a confidence value to incorporate our finding that the relative pressure to speak is time-sensitive within the first second of a gap (as shown in Figure 3). Variations in this value indicate the importance of speaking at a specific time.

Finally, a minimal turn-taking timing module would receive IUs where the payload may be an input signal (e.g., the audio signal). Additionally, it requires the ability to determine the elapsed time between turns in the conversation up to that point. Therefore, turn-taking module IUs will use the grounding-in relation to determine the specific IUs the results are based on.

4.3 Turn-Taking Timing Module

The turn-taking module consumes LB IUs to produce RB-IUs, with the invariant $size(RB) \leq size(LB)$, and implements the purge, new IU update, and commit operators. Here, the purge operator is vital in removing all IUs when a connected module, such as the ASR module, indicates that an interlocutor has taken the floor. In this case, any considerations for the time after which the system may start a turn depends only on the following turn and previous results may be discarded. The processor also implements the new IU update method to modify its internal state based exclusively on new LB IUs. It may then produce new turn-timing decisions based on the updated information. The commit method then finalizes the best-guess time

after which a turn may be started by the spoken dialogue system.

4.4 Module Extensions

In this section, we proposed a *minimal* incremental module for timing turn-taking based primarily on TRP duration. However, there are additional sources of information, not considered in this paper, that a spoken dialogue system may use when deciding when to produce a turn after a TCU. For example, intonational and semantic end of turn cues can be used to predict when floor transfer may occur (Lala et al., 2019; de Ruiter, 2019), and non-verbal cues may also be used to time turn-taking (de Ruiter et al., 2006a; Duncan, 1972). These may be modeled as individual IMs in the framework we use and connected to the turn-taking IM to allow information to be integrated when making turn-taking decisions. Additionally, the module may remember wait times proposed across a conversation and adjust for the specific interlocutor. For example, if there is frequent overlap if speech is produced after the proposed wait time, then future wait time estimates may be corrected. Similarly, short gaps by the interlocutor may be mimicked by the spoken dialogue system.

5 Conclusion and Future Work

In this study, we started by comparing two conceptual models of turn-taking—Duncan's "turn-yielding" cue model and Sacks et al.'s "simplest systematics", each of which makes different predictions about the organization of turns in conversation. We used data from the Switchboard corpus to fit two probabilistic models of TRP duration based on these conceptual models: a speaker-agnostic model compatible with Duncan's conceptual model and a speaker-sensitive model inspired by on Sacks' et al.'s conceptual model. Both models have a good quantitative and qualitative fit with the empirical data.

However, when comparing the two models directly, we found that the speaker-sensitive model i.e., Sacks et al.'s model, was decisively better at predicting the data than the speaker-agnostic model. We explored the implications of this finding for turn-taking systems. We showed that the likelihood of a speaker beginning a TCU during a pause in conversation changes as the pause lengthens. For short pauses, it is more probable that the speaker will switch, but as the pause continues, the original speaker becomes more likely to continue their turn.

Our work supports the notion that, for proper turn-taking, detecting and/or anticipating the end of turns is not sufficient. People are sensitive to the pauses and gaps in conversation and organize their speech to take into account this paralinguistic signal. We described the regularities that we found, and outlined implementations for dialogue systems to incorporate our findings. For naturalistic turn-taking adhering to these subtle norms is important, and we described first steps towards implementing this in agents.

In future work, we plan on implementing the spoken dialogue system we have proposed in this paper. While we have established and operationalized normative turn-taking behavior based on human conversations, it is important to investigate whether and to what degree findings from human-human data generalize to communication with spoken dialogue systems. Therefore, evaluating the conversational *naturalness* of our system through human-subject experiments is a relevant next step and will provide insight into the organization of turns in conversation, both for human-human and human-agent communication.

Acknowledgments

This paper was funded in part by a grant from the Data Intensive Studies Center at Tufts University.

References

Dan Bohus and Eric Horvitz. 2010. Computational models for multiparty turn taking. *Technical Report. Microsoft Research Technical Report MSR-TR 2010-115*.

Louis ten Bosch, Nelleke Oostdijk, and Jan P. de Ruiter. 2004. Durational aspects of turn-taking in spontaneous face-to-face and telephone dialogues. In *International Conference on Text, Speech and Dialogue*, pages 563–570. Springer.

Sara Bögels and Francisco Torreira. 2015a. Listeners use intonational phrase boundaries to project turn ends in spoken interaction. *Journal of Phonetics*, 52:46–57.

Sara Bögels and Francisco Torreira. 2015b. Listeners use intonational phrase boundaries to project turn ends in spoken interaction. *Journal of Phonetics*, 52:46–57.

Starkey Duncan. 1972. Some signals and rules for taking speaking turns in conversations. *Journal of Personality and Social Psychology*, 23(2):283–292.

Jens Edlund, Joakim Gustafson, Mattias Heldner, and Anna Hjalmarsson. 2008. Towards human-like spoken dialogue systems. *Speech communication*, 50(8-9):630–645.

Cecilia E Ford and Sandra A Thompson. 1996. intonational, and pragmatic resources for the. *Interaction and grammar*, 13:134.

Felix Gervits, Ravenna Thielstrom, Antonio Roque, and Matthias Scheutz. 2020. It's about time: Turn-entry timing for situated human-robot dialogue. In *Proceedings of the Special Interest Group on Discourse and Dialogue*.

John Godfrey and Edward Holliman. 1993. Switchboard-1 release 2 ldc97s62. *Linguistic Data Consortium*.

B. Goodrich, J. Gabry, I Ali, and S. Brilleman. 2020. Rstanarm: Bayesian applied regression modeling via stan r package v. 2.19.2.

Agustín Gravano and Julia Hirschberg. 2011. Turn-taking cues in task-oriented dialogue. *Computer Speech & Language*, 25(3):601–634.

Quentin F. Gronau, Henrik Singmann, and Eric-Jan Wagenmakers. 2020. bridgesampling: An r package for estimating normalizing constants. *Journal of Statistical Software*, 92(10).

Kohei Hara, Koji Inoue, Katsuya Takanashi, and Tatsuya Kawahara. 2019. Turn-taking prediction based on detection of transition relevance place. In *INTERSPEECH*, pages 4170–4174.

Mattias Heldner and Jens Edlund. 2010. Pauses, gaps and overlaps in conversations. *Journal of Phonetics*, 38(4):555–568.

Gail Jefferson. 1983. Notes on a possible metric which provides for a standard maximum silence of approximately one second in conversation. *Tilburg papers in language and literature*.

Kristiina Jokinen, Hirohisa Furukawa, Masafumi Nishida, and Seiichi Yamamoto. 2013. Gaze and turn-taking behavior in casual conversational interactions. *ACM Transactions on Interactive Intelligent Systems (TiiS)*, 3(2):1–30.

Casey Kennington, Daniele Moro, Lucas Marchand, Jake Carns, and David McNeill. 2020. rrsds: Towards a robot-ready spoken dialogue system. In *Proceedings of the 21th annual meeting of the special interest group on discourse and dialogue*, pages 132–135.

Divesh Lala, Koji Inoue, and Tatsuya Kawahara. 2019. Smooth turn-taking by a robot using an online continuous model to generate turn-taking cues. In *2019 International Conference on Multimodal Interaction*, pages 226–234.

Nathan P Lemoine. 2019. Moving beyond noninformative priors: why and how to choose weakly informative priors in bayesian analyses. *Oikos*, 128(7):912–928.

Stephen C Levinson and Francisco Torreira. 2015. Timing in turn-taking and its implications for processing models of language. *Frontiers in psychology*, 6:731.

Anthony J Liddicoat. 2004. The projectability of turn constructional units and the role of prediction in listening. *Discourse Studies*, 6(4):449–469.

Ryo Masumura, Mana Ihori, Tomohiro Tanaka, Atsushi Ando, Ryo Ishii, Takanobu Oba, and Ryuichiro Higashinaka. 2019. Improving speech-based end-of-turn detection via cross-modal representation learning with punctuated text data. In *2019 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*, pages 1062–1069. IEEE.

Ryo Masumura, Tomohiro Tanaka, Atsushi Ando, Ryo Ishii, Ryuichiro Higashinaka, and Yushi Aono. 2018. Neural dialogue context online end-of-turn detection. In *Proceedings of the 19th Annual SIGdial Meeting on Discourse and Dialogue*, pages 224–228.

Thilo Michael. 2020. Retico: An incremental framework for spoken dialogue systems. In *Proceedings of the 21th Annual Meeting of the Special Interest Group on Discourse and Dialogue*, pages 49–52.

Carina Riest, Annett B Jorschick, and Jan P de Ruiter. 2015. Anticipation in turn-taking: mechanisms and information sources. *Frontiers in psychology*, 6:89.

Felicia Roberts, Alexander L. Francis, and Melanie Morgan. 2006. The interaction of inter-turn silence with prosodic cues in listener perceptions of "trouble" in conversation. *Speech Communication*, 48(9):1079–1093.

Jan P. de Ruiter. 2019. Turn-taking. *The Oxford Handbook of Experimental Semantics and Pragmatics*, page 536–548.

Jan P. de Ruiter, H. Mitterer, and N. J. Enfield. 2006a. Projecting the end of a speaker's turn: A cognitive cornerstone of conversation. *Language*, 82(3):515–535.

Jan P. de Ruiter, Holger Mitterer, and Nick J Enfield. 2006b. Projecting the end of a speaker's turn: A cognitive cornerstone of conversation. *Language*, 82(3):515–535.

Harvey Sacks, Emanuel A. Schegloff, and Gail Jefferson. 1974. A simplest systematics for the organization of turn-taking for conversation. *Language*, 50(4):696–735.

David Schlangen and Gabriel Skantze. 2011. A general, abstract model of incremental dialogue processing. *Dialogue & Discourse*, 2(1):83–111.

Margret Selting. 2000. The construction of units in conversational talk. *Language in society*, 29(4):477–517.

Gabriel Skantze. 2021. Turn-taking in conversational systems and human-robot interaction: a review. *Computer Speech & Language*, 67:101178.

Gabriel Skantze and Anna Hjalmarsson. 2010. Towards incremental speech generation in dialogue systems. In *Proceedings of the SIGDIAL 2010 Conference*, pages 1–8.

Gabriel Skantze and David Schlangen. 2009. Incremental dialogue processing in a micro-domain. In *Proceedings of the 12th Conference of the European Chapter of the ACL (EACL 2009)*, pages 745–753, Athens, Greece. Association for Computational Linguistics.

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

Tanya Stivers and Jeffrey D Robinson. 2006. A preference for progressivity in interaction. *Language in society*, 35(3):367–392.

Ruud Wetzels, Dora Matzke, Michael D Lee, Jeffrey N Rouder, Geoffrey J Iverson, and Eric-Jan Wagenmakers. 2011. Statistical evidence in experimental psychology: An empirical comparison using 855 t tests. *Perspectives on Psychological Science*, 6(3):291–298.

Ran Zhao, Oscar J Romero, and Alex Rudnicky. 2018. Sogo: a social intelligent negotiation dialogue system. In *Proceedings of the 18th International Conference on intelligent virtual agents*, pages 239–246.

A Appendix

A.1 Statistical Model Priors and Parameters

The statistics in the table describe the priors used to fit the truncated normal distributions for the models described in the turn-taking models Sections 3.2 and 3.3. For each model, identical priors were used so that differences between the models were functions of the data, not of the priors. μ was drawn from a normal distribution with priors shown, and σ was drawn from a Gamma distribution with priors shown. Gamma distributions are typically parameterized with α and β parameters, but `pymc3` allows for parameterization with μ and σ, which is what we chose. All models took 10,000 samples with 6,000 tuning steps and a target acceptance rate of 0.9.

μ_μ	200
μ_σ	75
σ_μ	300
σ_σ	200

Table 3: This table shows the prior parameters used in each of the statistical models.

A.2 Model Comparison Methods

To compare the two models in Section 3.4, while taking into account model complexity, we built two linear mixed effects models using the `stan_glmer` function in the `rstanarm` package for the `R` programming language. This function fits a linear model of the data based on the parameters involved. Both models corrected for the particular conversation as a random effect, and one took into account whether there was a speaker switch at the TRP. Unlike `pymc3`, `Stan` does not require the user to specify priors, but assigns weakly informative default priors based on the data. Using the `Stan` models allowed use the `bayesfactor_models` function of the `bayestestR` package to compare the models and determine if the speaker switch model better explained the data than the no speaker switch model.

A.3 Generalized Models

We recognize that truncated normal models may not be the most robust method of modeling our data - which does not include gaps and overlaps. Additionally, the truncated normal distribution used for the speaker continuation condition is only positive valued. Therefore, we present preliminary analyses on alternative models that may be used to fit our data.

The analyses presented in the paper establish that the speaker switch and continuation conditions are different and provide a justification for creating stochastic models to describe these phenomenon separately. Therefore, we built a Student's t-distribution model for the speaker switch condition to approximate the normal model when ν is large. The dataset used for this model includes all TRPs with duration in range -800 ms to +2500 ms, and only excludes outliers that were likely to be transcription artifacts. The widely-applicable information criterion (WAIC) score for the Student's t is 79,131, while the truncated normal (expanded to the new lower and upper limits) is 79,707, showing some improvement. The Kolmogorov-Smirnov

statistic is also reduced to 0.425 from 0.741 (both $p < 0.001$).

	$\mu_{\text{hdi 3\%}}$	μ_{model}	$\mu_{\text{hdi 97\%}}$	σ_{model}
ν	2.70	2.96	3.23	0.142
μ	95	105	116	5.50
σ	287	297	307	5.63

Table 4: This table describes the posterior model parameters used for the Student's t-distribution model in the speaker switch condition.

Additionally, we built a Gamma model for the speaker continuation condition using TRPs with duration up to 2500 ms. This model describes a variable with a positive domain more elegantly than a truncated normal model. The WAIC score of the Gamma and truncated normal (with adjusted bounds) models is 33,845 and 34,125 respectively, which again shows improvement. The Kolmogorov-Smirnov statistic is also reduced from 0.480 to 0.173 (both $p < 0.001$).

	$\mu_{\text{hdi 3\%}}$	μ_{model}	$\mu_{\text{hdi 97\%}}$	σ_{model}
μ	604	621	638	9.00
σ	419	435	451	8.64
α	1.95	2.06	2.16	5.71e-2
β	3.10e-3	3.29e-3	3.49e-3	1.04e-4

Table 5: This table describes the model parameters used for the Gamma model for the speaker continuation condition.

Each of the models presented show potential next steps for improving modeling our data. Unfortunately, for our purposes, they are not directly comparable.

A.4 Data Description

The truncated normal models described in Section 3 exclude some data – which was necessary for our analysis. Here, we include descriptions of our raw data to provide further information on our analyses.

Duration	Number of TRPs
0 ms	3565
0–1000 ms	1933
$>$ 1000 ms	380

Table 6: This table shows the number of speaker-continuation TRPs in bins of different duration.

The above descriptions show that overlapping speech is extremely common in our dataset, making up about 42% of the speaker switch conditions.

Duration	Number of TRPs
$<$ 0 ms	2217
0–1000 ms	2703
$>$ 1000 ms	296

Table 7: This table shows the number of speaker-switch TRPs in bins of different duration.

Additionally, though we only consider positive values, speaker continuations with pauses of 0 ms are the majority of speaker continuation conditions —61%. The models we have presented model reasoning *through* a silence, and are therefore sound in the assumption that a silence exists. However, any turn taking model that *only* considers turn taking via silences will be incomplete.