

# Using Past Speaker Behavior to Better Predict Turn Transitions

Tomer M. Meshorer

B.A. Computer Science, Tel Aviv University, 1997

Presented to the  
Center for Spoken Language Understanding  
within the Oregon Health & Science University  
School of Medicine  
in partial fulfillment of  
the requirements for the degree  
Master of Science  
in  
Computer Science & Engineering

March 2017

Copyright © 2016 Tomer M. Meshorer  
All rights reserved

Center for Spoken Language Understanding  
School of Medicine  
Oregon Health & Science University

---

CERTIFICATE OF APPROVAL

---

This is to certify that the M.S. thesis of  
Tomer M. Meshorer  
has been approved.

---

Prof. Peter A. Heeman, Thesis Advisor  
Research Associate Professor

---

Steven Bedrick  
Assistant Professor

---

Stephen Wu  
Assistant Professor

# Dedication

I would like to dedicate this work to my father, Israel Sagie (1945 – 2016), who passed away during the preparation of this work. His wisdom and support will always guide me.

# Acknowledgements

I would like to express my deep appreciation to Prof. Peter Heeman for his guidance during this research. Without his valuable assistance, I could not have completed this work.

I would also like to express my appreciation to the members of the committee, Prof. Steven Bedrick and Prof. Stephen Wu, for reviewing the work and providing invaluable feedback.

I am also indebted to the staff and students of the Center of Spoken Language Understanding, and especially to the graduate program coordinator, Ms. Patricia Dickerson, for her valuable assistance during the course of my studies. This work was partially funded by the National Science Foundation under grant IIS-1321146.

# Contents

<b>Dedication</b> . . . . .	<b>iv</b>
<b>Acknowledgements</b> . . . . .	<b>v</b>
<b>Abstract</b> . . . . .	<b>viii</b>
<b>I Introduction</b>	<b>1</b>
<b>1 Introduction</b> . . . . .	<b>2</b>
1.1 Thesis Statement . . . . .	3
1.2 Approach . . . . .	3
1.3 Contribution . . . . .	4
1.4 Dissertation Structure . . . . .	4
<b>2 Related Work</b> . . . . .	<b>5</b>
2.1 Human-Human Conversations . . . . .	5
2.2 Human-Computer Conversations . . . . .	6
<b>3 Theoretical model</b> . . . . .	<b>8</b>
3.1 Local Features . . . . .	8
3.2 Summary Features . . . . .	8
<b>4 Data</b> . . . . .	<b>10</b>
4.1 Data Preparation . . . . .	10
4.2 Data Exploration . . . . .	12
4.2.1 Dialog Acts . . . . .	12
<b>5 Study</b> . . . . .	<b>17</b>
5.1 Classification Models . . . . .	17
5.2 Metrics . . . . .	17
5.3 Results . . . . .	19
5.3.1 Sensitivity to measurement start time . . . . .	20
<b>6 Conclusions</b> . . . . .	<b>22</b>
6.1 Future Direction . . . . .	22

<b>Bibliography</b> . . . . .	<b>24</b>
-------------------------------	-----------

# Abstract

## Using Past Speaker Behavior to Better Predict Turn Transitions

Tomer M. Meshorer

Master of Science

Center for Spoken Language Understanding  
within the Oregon Health & Science University  
School of Medicine

March 2017

Thesis Advisor: Prof. Peter A. Heeman

Conversations are at the core of everyday social interactions. The interactions between conversants are preformed within the realm of a sophisticated and self-managed turn taking system. In human conversations, the turn taking system supports minimal speaker overlap during turn transitions and minimum gaps between turns. Spoken dialogue systems are a new form of conversational user interface that permits users to use their voice to interact with the computer. As such, the turn taking capabilities of SDS should evolve from a simple timeout to a more human-like model. Recent advances in turn taking systems for SDS use different local features of the last few utterances to predict turn transition.

This thesis explores using a summary of past speaker behavior to better predict turn transitions. We believe that the summary features represent an evolving model of the other conversant. For example, speakers who typically use long turns will be likely to use long turns in the future. In addition, speakers with more control of the conversation floor will be less likely to yield the turn. As the conversational image of the speaker evolves as the conversation progresses, other speakers might adjust their turn taking behavior in response.

We computed two types of summary features that represent the current speaker's past turn-taking behavior: *relative turn length* and *relative floor control*. Relative turn length measures the current turn length so far (in seconds and words) relative to the speaker's average turn length.



Relative floor control measures the speaker’s control of the conversation floor (in seconds and words) relative to the total conversation length. The features are recomputed for each dialog act based on past turns of the speaker within the current conversation. Using the switchboard corpus, we trained two models to predict turn transitions: one with just local features (e.g., current speech act, previous speech act) and one that added the summary features. Our results shows that using the summary features improves turn transition prediction.

## **Part I**

# **Introduction**

# Chapter 1

## Introduction

Conversations are a frequent form of everyday social interaction and are characterized by a rapid exchange of messages between the conversants. The turn exchange system in human conversation is universal in nature and crosses culture, age and language [20]. According to the seminal work by Sack et al. [34], only one conversant is usually speaking at any given time, conversants takes turns, and the gap and overlap between turns is kept to a minimum. These attributes apply regardless of turn length (from a single word to a full sentence) and conversation length.

In this work we are interested in the point of possible transition between speakers. To keep the gaps between turns minimal while supporting speaker change, the listener must predict a possible turn transition before the end of the speaker's utterance. Two dominant approaches tried to explain the mechanisms by which the conversation floor is allocated. Duncan [6] suggested that the speaker signals the listener about an upcoming turn transition by using a combination of one or more signals. Another approach, suggested by Sacks et al. [34], defines the turn allocation process in terms of a set of local rules that are applied at possible turn transition points. Both approaches (signaling and turn allocation rules) are based on phenomena that occur in the last few utterances (for example, the syntactic construct of the turn or the use of an adjacency pair<sup>1</sup>). This work investigates whether using features that are computed over all past turns can help improve the predictability of turn transitions.

Spoken dialogue systems (SDS) are computer systems that support a conversational speech-based user interface. Users engage in a conversation with the computer to perform a task (for example, information seeking) by using their natural apparatus, the voice. Hence, to be effective and user friendly, an SDS should also adhere to the system of turn exchanges.

Early SDS systems did not contain a turn management component and instead used a fixed

---

<sup>1</sup>Adjacency pairs are a pair of utterances which have the following attributes [36]: (1) adjacent positioning of component utterances, (2) different speakers producing each utterance, (3) the first utterance relates in a specific way to the second utterance thus forming a pair type, for example, Question-Answer.

timeout to detect the end of a user turn. Using a simple timeout led to barge-in situations, in which the system prematurely started to speak during the user’s turn. Decreasing the barge-ins by increasing the timeout caused large gaps between turns, where the user waited for the system to speak. To improve this situation, newer SDS systems are incorporating recent findings in human-human conversation in their prediction models. For example, features from the latest utterance are used to predict turn transition. Prediction in human-human conversation is based on syntactic [34, 5], prosodic [10, 7, 8], and pragmatic cues [9].

While local features of the latest utterance form an important input for prediction, this thesis postulates that speakers might also use summary features computed over many turns. The summary features form a conversational image of the speaker and contain features that represent the speaker’s average behavior over many turns. For example, average turn length measures the length in both time and words of each conversants turn up to this turn. Hence, if the length of the current speaker’s turn is more than its average turn length, it is more likely that a turn ending will occur.

## 1.1 Thesis Statement

This work investigates whether summary features based on a speaker’s past behavior in a conversation can help improve turn transition prediction. Using the switchboard corpus, we computed for every dialog act in the conversation two summary features: *relative turn length* of the current turn, and the *relative floor control*. In addition, after each dialog act in the conversation, we predict whether a turn transition occurred. Our hypothesis is that we can make better turn transitions by using the summary features compared to using only local features.

## 1.2 Approach

To test the effectiveness of the summary features, we used the NXT version of the Switchboard corpus [4, 12] to train random forest models [27]. First we performed data preprocessing in order to merge turn, dialog act and single word information into a single data set. Next, we trained baseline models using local features: current and previous dialog acts. We also trained a model on the summary features as well as a model that includes both the local and the summary features. To measure the predictive power of the different models we use standard model metrics: F1, precision, recall and area under the curve (AUC)

### 1.3 Contribution

Our results show that the summary features improve prediction performance in both area under the curve (AUC), 69.4% vs 66%, and F1, 69.3% vs 57.8%. In addition, the model that was trained on all of the features performs better than the local features model in both AUC, 83.7% vs 81.1%, and F1, 77.5% vs 74.8%. The results show that using summary features helps predict turn transitions.

### 1.4 Dissertation Structure

The rest of this thesis is as follows. Chapter 2 presents the related work in both human-human and human-computer conversations. Chapter 3 describes the theoretical model that underlies the summary features. Chapter 4 describes the data sources, data preparation and initial data analysis. Chapter 5 describes our main study in which we compute the long term features over all the turns in each conversations. The chapter compares the baseline models as well as the model that includes the summary features. Finally, in chapter 6 we present the conclusion and future work.

This thesis is based on a paper [26] that was submitted and presented at Interspeech.

# Chapter 2

## Related Work

This chapter presents work related to turn transition prediction in both human-human conversations and human-computer conversations. Studying human-human turn allocation mechanism might give us a clue on the inner working that humans use during conversations when regulating turn allocation. This in turn gives us a clue on how to implement such mechanisms in spoken dialogue systems.

### 2.1 Human-Human Conversations

Duncan [6] argued that speakers signal when they want the listener to take the turn and presented six signals used by the speaker to accomplish this: intonation, drawl on the final syllable, body motion, sociocentric sequence, drop in pitch or loudness, and syntax. Kendon [18] included a gaze as an additional signal for turn transition. As the signals are local in nature (derived from the last turn), our summary features complement them.

Turn allocation was introduced in the seminal work by Sacks, Schegloff, and Jefferson [34], who observed that conversations are “one speaker at a time” and gaps between turns as well as speaker overlaps are kept to a minimum. To satisfy these constraints, Sacks et al. suggested an ordered set of rules for turn allocation: (a) current speaker selects the next conversant; (b) if the current speaker did not select, any of the listeners can self select; or (c) if neither of the previous two cases apply, the current speaker continues. For the first rule, Sacks et al. suggested that the current speaker uses adjacency pairs as the main apparatus for selecting the next speaker. Hence, we recognized the importance of dialog acts in turn allocation and chose them as the atomic turn components. In addition, our work might impact the second rule, in which the conversant self selects. While Sacks et al. suggested that the first starter is the next speaker, we suggest that a conversant might use the conversational image of the speaker and of themselves when self selecting. For example, a controlling speaker (with a high relative floor control score) has a better chance to

gain control of the conversation floor when self selecting.

In addition to the turn allocation system, Sacks et al. also suggested that turn construction units (TCU) should support projection of turn ends by the participants. The projectability attribute was later extended to other features of the speaker’s utterance: (syntactic [34], prosodic [10] and pragmatic [10, 9]). Our work augments the local utterance features with summary features that can be used to improve projectability.

Selfridge and Heeman [38] work on turn bidding suggests that each conversant measures the importance of their utterance when negotiating the right to the conversation floor. Their work is based on the fact that 1) The importance of speaking is the primary motivation behind turn-taking behavior, and (2) conversants use turn cue strength to bid for the turn based on this importance. Our work suggest that contestants might use summary feature to cue the strength of the bid. For example conversant will take turn longer than their average if they want to signal turn importance.

Wilson et al [45] explored the role of entrainment in turn taking. They proposed there is entrainment of endogenous oscillators in the brains of the speaker and the listener on the basis of the speaker syllabus production. In their model, the speaker and the listener are counter-phased so that speech overlaps and gaps are minimized. Although our work does not imply cyclic synchronization between speaker and listener, we do suggest that each conversant creates a conversation image of the other conversant and uses it during turn transition.

The importance of using dialog acts was emphasized by a very recent study of Garrod and Pickering [11]. The study suggested that turn production is a multi-stage process in which the listener performs simultaneous comprehension of the existing turn as well as production of the new turn content. They suggested that the first step in the process is dialog act recognition, which is done as soon as possible and acts as the basis for the listener’s turn articulation and production. In our study we use dialog act as the main turn component.

## 2.2 Human-Computer Conversations

As recent advances in machine learning reduce speech recognition error rates, the problem of turn taking in SDS rises in importance [17]. Traditional SDS systems use a simple silence timeout approach to trigger turn transitions. This creates three issues [1] first, the model might not be robust enough in a noisy environment (for example when driving); second, if the timeout is too short, the system might detect intra-turn pauses (for example, the user pausing to think) as a turn transition and will cut into the user’s turn; and third, if the timeout is too long the system will wait too long to take the turn, resulting in large gaps between turns.

Recent studies tried to improve over the simple threshold model by using machine learning to train models based on features derived from the latest utterance. As different studies use a variety of features, we will focus on those that are similar to our summary features.

Gravano and Hirschberg [14] used the Columbia games corpus to study the effectiveness of different turn transition cues. The authors define inter-pausal units (IPU) as a maximum sequence of words surrounded by silence of more than 50 ms. A turn is the longest sequence of IPUs by the same speaker. One of the features studied is IPU duration in ms as well as number of words. As in our findings, the authors found that long IPUs are a good indication of upcoming turn changes (long IPUs might correlate with a speaker passing its average turn length). Moreover, as we show in Section 5, the authors found that combining multiple cues leads to better accuracy.

Raux and Eskenazi [30] performed a comprehensive study of features that inform turn changes. The study found that timing features, such as turn duration and number of pauses, have relatively strong predictive power. While Raux and Eskenazi use features of the current turn, in our study we use the timing features for the turns that have occurred so far in the current conversation.

In more recent study, Nishitha and Rodney [15] used a model based on N grams of dialog acts to predict turn transitions. They trained decision tree models using the switchboard data. As features, they contrasted the use of the previous dialog act, the previous two dialog acts and the previous three dialog acts. They also contrasted whether to indicate if there is a speaker change between the previous dialog acts. They achieved an F1 measure of 0.67 for the model that use the previous two dialog acts. In this paper, we based our baseline models on previous dialog act and the previous two dialog acts. We also mapped the switchboard dialog acts from 148 dialog acts down to 9 to reduce data dimensionality. The prediction performance of our baseline model is comparable to their results.



# Chapter 3

## Theoretical model

This chapter aims to provide a more formal description of our turn taking model. First we present a formal definition of a conversation, turn changes and local features. Next we presents the formulas to compute the summary features: relative turn length and relative floor control.

### 3.1 Local Features

We define a conversation as a sequence of dialogue acts  $d_1 \dots d_N$ , where  $d_i$  is uttered by speaker  $s_i$ . We write this as the following sequence:

$$\dots s_{i-2}, d_{i-2}, s_{i-1}, d_{i-1}, s_i, d_i \dots \quad (3.1)$$

We denote whether there was a turn transition with  $y_i$ . A turn transition occurs when the speaker  $s_i$  is different from speaker  $s_{i-1}$ . Hence, (1) can be also be viewed as a sequence of dialog acts  $d_i$  followed by turn transitions  $y_i$ :

$$\dots d_{i-2}, y_{i-1}, d_{i-1}, y_i, d_i, y_{i+1} \dots \quad (3.2)$$

In our first baseline model, we try to predict the turn transition value  $y_{i+1}$  based only on the latest dialog act  $d_i$ . In our second baseline model, we try to predict turn transition  $y_{i+1}$  based on the latest two dialog acts:  $d_{i-1}$  and  $d_i$ .

### 3.2 Summary Features

As discussed in the introduction, we introduce two types of summary features in this paper: relative turn length ( $rt_i$ ) and relative floor control ( $rc_i$ ). These features are used in predicting whether there is a change in speaker  $y_{i+1}$  after dialogue act  $d_i$ .

To compute the summary features at dialogue act  $d_i$ , we introduced the some notation. Let  $S_i$  to be the set of complete turns of speaker  $s_i$  that are prior to the turn that  $d_i$  is in. Let  $t_i$

represent the turn so far that  $d_i$  is in, up to  $d_i$  but no subsequent dialogue acts. Let  $\text{length}(t)$  be the length of a turn or a partial turn in seconds (or words). To compute the *relative turn length* of turn  $t_i$  we first compute the average length of all the turns in  $S_i$

$$\text{avg\_}t_i = \frac{\sum_{t \in S_i} \text{length}(t)}{|S_i|} \quad (3.3)$$

The *relative turn length* summary feature of  $t_i$ , denoted as  $rt_i$ , measures the percent of the length of the turn  $t_i$  so far, relative to the average turn length up to  $t_i$  of the current speaker  $s_i$  (but not including  $t_i$ ).

$$rt_i = \frac{\text{length}(t_i)}{\text{avg\_}t_i} \quad (3.4)$$

Note that we calculate two versions of  $rt_i$ : in seconds and in words. The purpose of  $rt_i$  is to let the listener, in predicting turn changes, take into account whether the current speaker is exceeding his or her average turn length.

The *relative floor control*, denoted as  $rc_i$ , measures the percent of time in which the current speaker controlled the conversation floor up to  $d_i$ . We use  $S_i$  as defined above: the set of complete turns of speaker  $s_i$  that are prior to the turn that  $d_i$  is in. We similarly define  $L_i$  to be the turns of the other conversant (the listener of  $d_i$ ) that are prior to  $d_i$ . We first compute the conversation length up to  $d_i$  denoted as  $c_i$ , which excludes inter-utterance pauses but includes intra-utterance pauses.

$$c_i = \sum_{t \in S_i \cup L_i} \text{length}(t) \quad (3.5)$$

To compute relative floor control at  $d_i$ , we divide the floor time of the speaker  $s_i$  up to turn  $t_i$  by  $c_i$ :

$$rc_i = \frac{\sum_{t \in S_i} \text{length}(t)}{c_i} \quad (3.6)$$

Note that we calculate  $rc_i$  in seconds and in words. Participants can use the relative floor control as a means to determine if one speaker is controlling the conversation; a controlling speaker will probably be less inclined to give up the floor.

We use these two summary features in the *summary model* and *full model*, as described in the next section.

# Chapter 4

## Data

To evaluate the importance of the summary features in predicting turn transitions we used the 2010 version of the switchboard corpus [4] which is based on the original release [13].

The switchboard corpus is the first large collection of phone conversations and was collected in 1990-1991. The initial goal of the corpus is to facilitate speech research by providing pre recorded audio files of day to day conversations. The original corpus was composed of 2483 phone calls involving 520 speakers. Each call ranged from 1.5 minutes to 10 minutes, with an average length of 6 minutes. Conversations involved a randomly chosen topic between two randomly selected speakers. The original audio recordings were later annotated and was released as part of the Penn Treebank 3 corpus, which included 650 annotated conversations. The current release used for this research includes 642 conversations and just over 830000 words.

The current corpus contains multiple types for annotations for each conversation. We used the turns annotation to assign turns to speakers and mark the turn start and end points. We used the dialog act annotation to assign dialog acts to turns such that each turn interval contain all the dialog acts that occurred between turn start and turn end. Next, we use the token annotation to assign words to dialog acts and to compute the overall number of words within each dialog act.

### 4.1 Data Preparation

Figure 4.1 shows the experiment data pipeline. Data is imported from the NXT switchboard corpus [4] into a graph database [44]. Figure 4.2 shows the data structure as it is represented inside the graph database. For each conversation, the conversation entities (words, dialog acts and turns) are represented as edges between time points, which are represented as vertices. The structure leads to a direct computation of the summary features using the graph query language.

After computing the summary features, we perform the following data transformation:

- We exclude 11 dialogue acts that were coded in Switchboard as “other.”

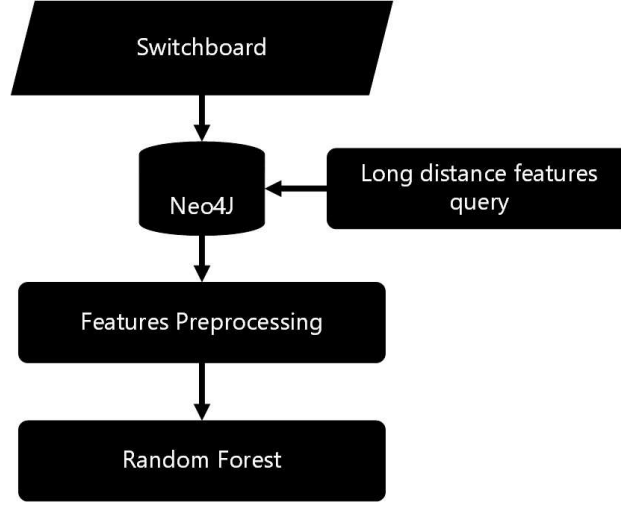


Figure 4.1: Experiment data pipeline

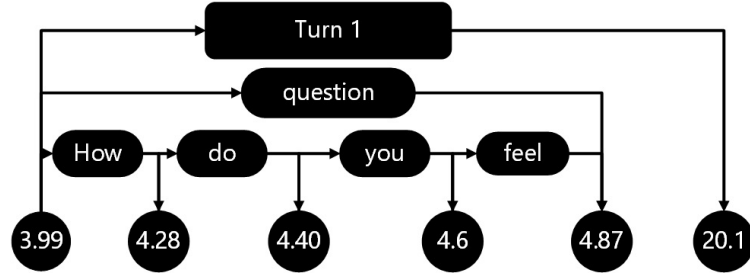


Figure 4.2: Conversation graph data model

- We filtered out all the dialogs that had data integrity issues, for example dialog acts that referred to non existent terminals. This decrease the number of conversations from 624 to 310 conversations consisting of 60595 dialog acts.
- To reduce data sparsity, we grouped switchboard dialog acts into dialog act classes. This reduced the number of dialog acts from 148 to 9 dialog act classes. See Table 1 for examples of the mapping.
- We added a binary  $y_{i+1}$  feature to each dialog act. As explained in Section 3, the variable is 1 if there is a turn change from dialogue act  $d_i$  to  $d_{i+1}$ .

Switchboard dialog acts	Dialog act classes
sd,h,bf	statement
sv,ad,sv@	statement - opinion
aa,aaâ	agree accept
%.%-%,%@	abandon
b,bh	backchannel
qy,qo,qh	question
no,ny,ng,arp	answer
+	+
o@,+@	NA

Table 4.1: Mapping from dialog act to dialog act class

## 4.2 Data Exploration

In this section we use visualization method in order to explore the distribution of the independent data variables and thier relationships with the turn taking (which is the depended variable) The visualization was done using python seaborn library which is based on matplotlib. The data set for the visualization contains a row for each dialog act. For each dialog act we measure its length, and the values of the summary features (RTL and RTC). The independent variable denote wether a turn change occured after the dialog act.

Variable	Description	Type
Previous Dialog Act	the dialog act before the current one	categorical
Dialog Act	the current dialog act	categorical
Length	length of the current dialog act in seconds	seconds
Relative Turn Length (RTL)	Relative turn length as defined in 3.2	percent
Relative Time Control (RTC)	Relative time control as defined in 3.2	percent
Turn Change	1 if there was a turn change after this dialog act	binary

Table 4.2: Data Fields

### 4.2.1 Dialog Acts

The first visualization, Figure 4.3 plots the distribution of each dialog act in the corpus. Each bar is a count of the dialog act type, divided by the total dialog acts. We can observe that the majority of dialog acts are statements, backchannels and opinions. This is true to the nature of the switchboard corpus which consists mainly of casual conversations.

Next we wanted to measure the correlation between dialog act type and the probability of a turn change. In figure 4.4, each bar measures the number of time a turn change occurred after

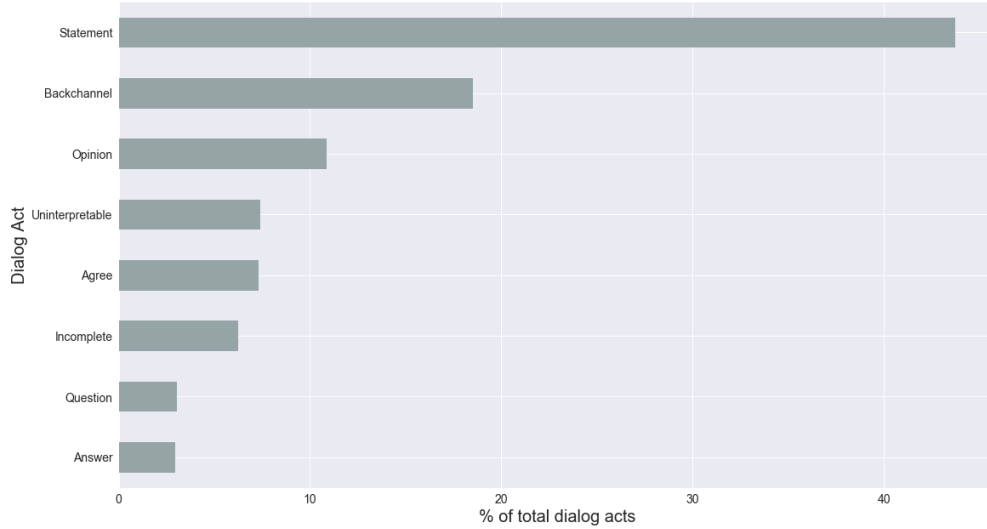


Figure 4.3: Dialog act relative count

a dialog act type divided by the total dialog act of this type. Or the probability that a dialog act of the said type will lead to turn change. We observe that the majority of back channels and question (which are usually the first utterance in a question-answer adjacency pair) will lead to a turn change.

Next we measured the relative turn length score for each type of dialog act. In 4.5, for each dialog act we show the distribution of relative turn length score for act that led to turn change and act that did not. We can see that in most of the dialog acts, the median RTL score that led to turn change is smaller than RTL score of acts of which the speaker kept the floor. This suggests that a speaker which uses the floor more than its average turn length will tend to hold the floor even more.

Next we measured the relative floor control score for each type of dialog act. In 4.8, for each dialog act we show the distribution of relative floor control for act that led to turn change and act that did not. We can see that in most of the dialog acts, the median score is about equal and close to 50%. We can observe that for the question dialog act, when the speaker had high relative floor control, it tends to keep the turn after the question (for example, by asking rhetorical question).

Next we measured the relation between values of relative turn length and probability of turn change. To compute figure 4.8, we divided the range of relative floor control into buckets and counted for each bucket the number of dialog acts (regardless of the type) that led to a turn

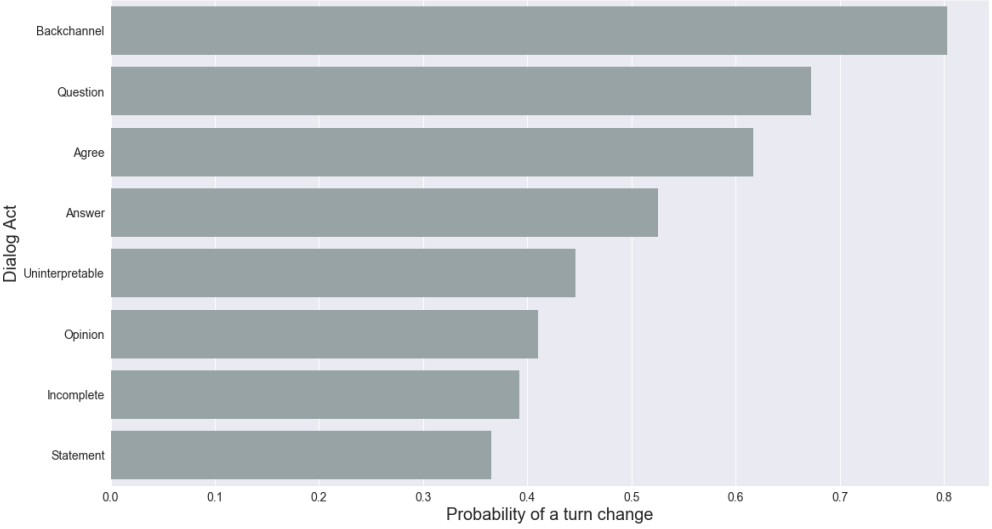


Figure 4.4: Dialog act probability of turn change

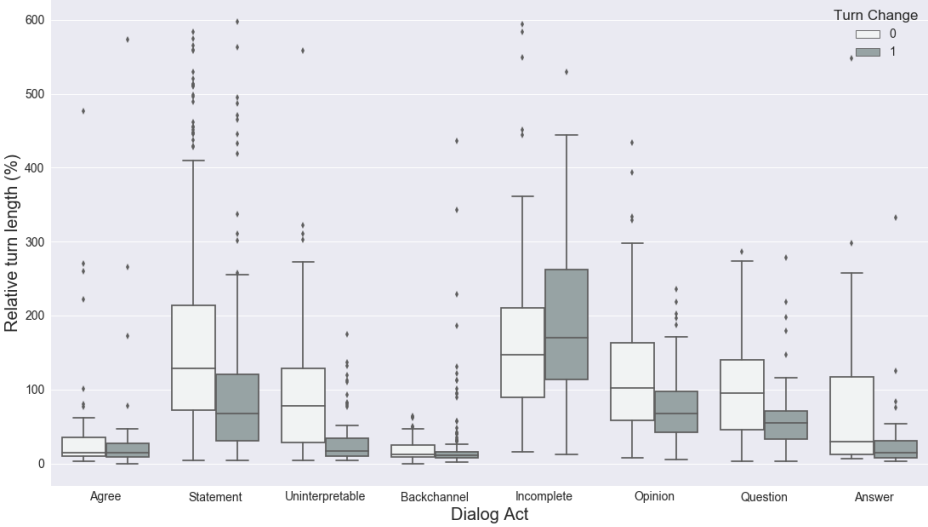


Figure 4.5: RTL score for dialog act type

change and the total dialog acts in the bucket. We then computed the probability of a turn change by dividing the former by the latter. As shown, when the chance of turn transfer is higher when

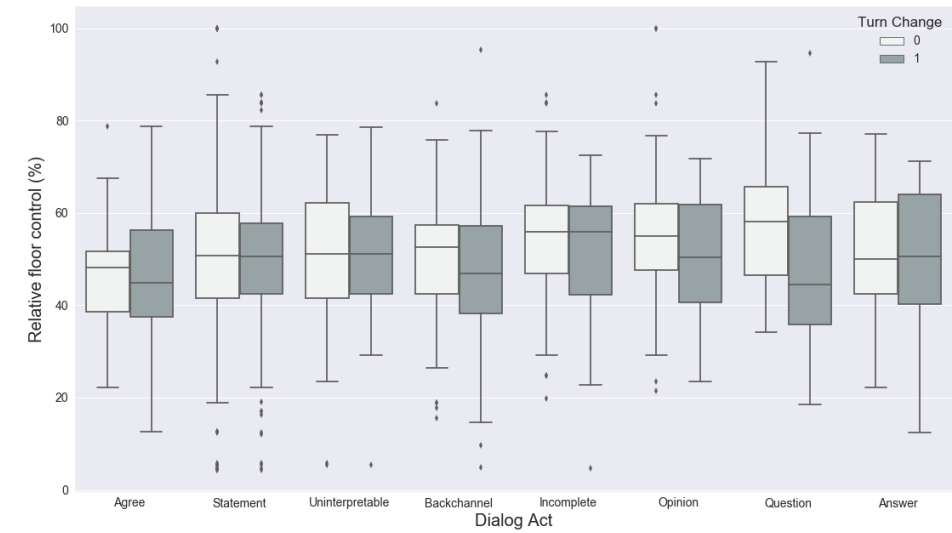


Figure 4.6: Dialog act probability of turn change

the speaker has the floor for the relative short amount of a time ( probably the affect of back channel). We can also observe that as the speaker has the floor for more time, it tend to hold it.

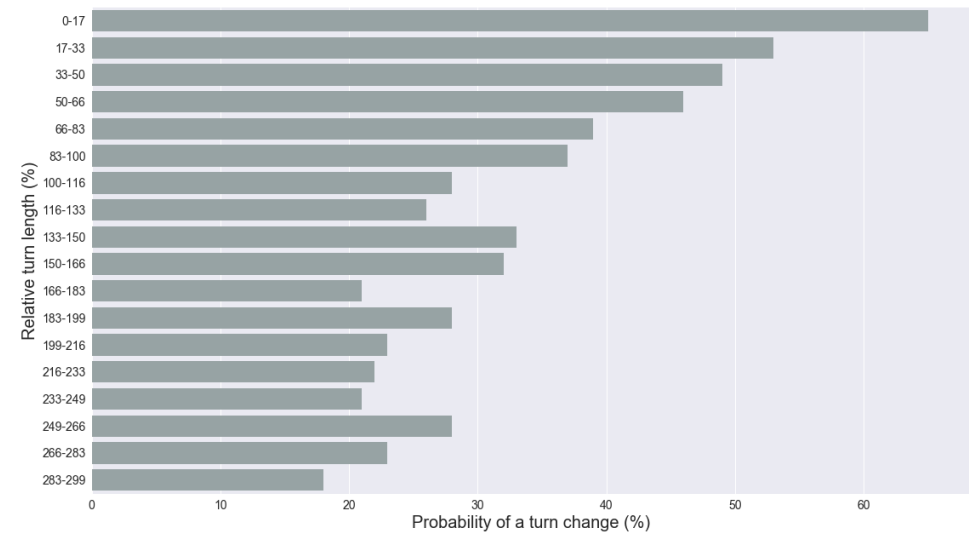


Figure 4.7: Relative floor control affect on probability of a turn change



Next we measured the relation between values of relative floor control and probability of turn change. As with relative turn length, to compute figure 4.8, we divided the range of relative floor control into buckets and counted for each bucket the number of dialog acts (regardless of the type) that led to a turn change and the total dialog acts in the bucket. Again, we then computed the probability of a turn change by dividing the former by the latter. As shown, we can see that the relative floor control does not have a lot of affect on the probability of a turn change. However, we can notice that when the speaker has relative floor control scores above 50% it tend to keep the floor and avoid turn change.

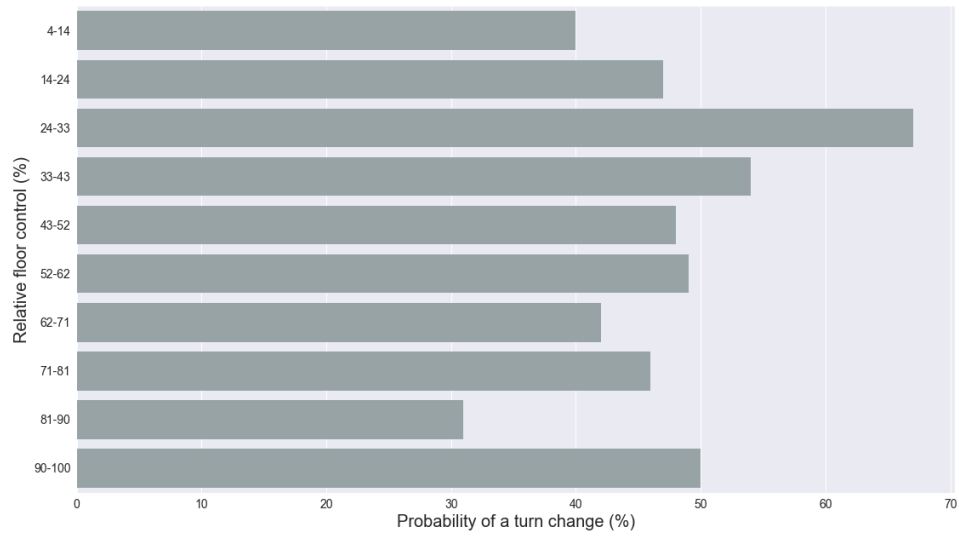


Figure 4.8: Dialog act probability of turn change

# Chapter 5

## Study

### 5.1 Classification Models

To test the contribution of the summary features, we used a binary classifier with  $y_i$  as the outcome variable. We trained four models, which used the following sets of features:

**baseline 1:** Predict turn transition based only on the current dialog act label.

**baseline 2:** Predict turn transition based on the labels of the current and previous dialog acts.

**summary model:** Predict turn transition using just the summary features.

**full model:** Predict turn transition using the summary features and the current and previous dialog acts.

We used random forests to build the binary classifiers ( $N = 200$ ) [3]. Random forests build an ensemble of decision trees during training, and during testing, each decision tree votes on the outcome. Like decision trees, they can account for interactions between variables, such as making greater use of the summary features when the current speech act is not a question. Random forests though are not as sensitive to overfitting and data fragmentation.

We performed 10 fold-labeled cross validations. We made sure that each conversation was entirely in a single fold. This way, each dialogue was entirely used for training or testing, but never for both at the same time.

### 5.2 Metrics

To evaluate our hypothesis, we use the trained model to perform prediction of the data. We then compare the predictions against the truth values. The results are shown in a confusion matrix as

		Predicted to cause turn change	
		yes	no
Caused turn change	yes	True Positive	False Negative
	no	False Positive	True Negative

Table 5.1: Confusion Matrix

shown in Table 5.1. Each row in the matrix represent the actual class and each column represent the predicted class. The cell in the matrix are compute as follow :

1. True Positive - Count of the dialog acts that were predicted to cause a turn change and that did lead to turn change.
2. False Positive- Count of the dialog acts that were predicted to cause a turn change but in fact did not cause one.
3. True Negative - Count of dialog acts that were predicted to not cause a turn change and did not cause turn change.
4. False Negative - Count of dialog acts that were predicted to not cause a turn change but in fact caused one.

Based on the confusion matrix, we compute the following metrics for each model:

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \quad (5.1)$$

The accuracy measures how many dialog acts were classified correctly out of the total dialog acts.

$$Precision = \frac{TP}{TP + FP} \quad (5.2)$$

The precision metric measures how accurate the classifier is for the the dialog acts that caused turn change; I.e., how many dialog acts that were classified as causing a turn change, did in fact led to a turn change.

$$+ - Recall = \frac{TP}{TP + FN} \quad (5.3)$$

The recall metric measure how many dialog act that predicted turn change were detected by the classifier.

$$F1 = \frac{2 \cdot Precision \cdot Recall}{Precision + Recall} \quad (5.4)$$

In order to have one measurement which encompass both recall and precision, we compute F1 which is the harmonic mean of recall and precision.

Note that there is a trade off between recall and precision: if we want to increase recall we will reduce precision. To measure this tradeoff, we will use ROC (Receiver Operating Characteristic) curve. The ROC curve measure the true positive rate (TPR) or recall, against the false positive rate (FPR). When comparing different ROC curves, we measures the area under the curve (AUC).

### 5.3 Results

We first analyzed the results in terms of accuracy: how often the models correctly predicted whether a turn transition occurred; in other words, how often the model predicts the correct value of  $y_{i+1}$ . Table 5.3 shows the results of training a random forest of 200 trees for each model using 10 folds cross validation. We see that using the summary features provides better accuracy than baseline 1, which use only the current dialog act (65.54% vs 62.79%). In addition, using the full model yields an improvement of over 1.08% in the accuracy. In addition, baseline 1 has high precision, it has very low recall. Using only the summary model improves recall and decreases precision by less, leading to a higher F1 score and overall better performance. Using the full model improves precision, which means that dialog acts that were considered to lead to turn transitions are classified correctly. If we use the full model, we lose precision (over baseline 2 model), but gain recall, leading to the highest F1 score and the best performance.

	Accuracy	F1	Precision	Recall	AUC
baseline 1	0.627938	0.578149	0.749824	0.470491	0.659992
baseline 2	0.748900	0.748792	0.818479	0.690012	0.811185
summary	0.655471	0.693203	0.672255	0.713612	0.694657
full	0.757515	0.775993	0.775099	0.778319	0.837863

Table 5.2: Precision, recall and F1 results using Random Forests

The effect can also be seen in figure 5.1, which shows the ROC curves and the AUC for each model. We notice that the AUC of the summary model is better than baseline 1 model (0.69 vs 0.66), and when adding the summary features to the local features (the full model), we see the

AUC improves (0.84 vs 0.81). This suggests that while the discrimination facility of the summary features is lacking ( $\text{AUC} < 0.7$ ), adding them to a classifier that uses local features (full model) yields better results than the baseline.

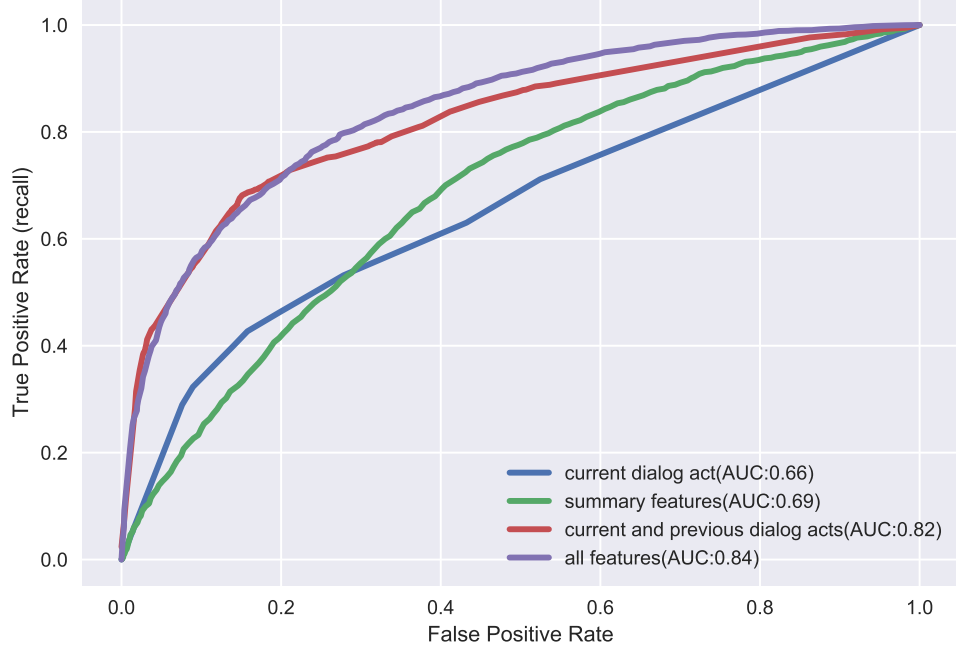


Figure 5.1: ROC curves and AUC of the different models

	accuracy	f1	precision	recall	roc_auc
all	0.765718	0.787481	0.774492	0.801152	0.848495
baseline 1	0.627938	0.578149	0.749824	0.470491	0.659992
baseline 2	0.748834	0.748210	0.819280	0.688687	0.811097
summary	0.679110	0.713018	0.692053	0.735588	0.726475

Table 5.3: Precision, recall and F1 results using Random Forests

### 5.3.1 Sensitivity to measurement start time

In Meshorer and Heeman [26] we assumed that it take 120 seconds for the conversational image to form. To validate this assumption here, we computed the AUC for the different model using different starting time. Table 5.4 shows that the AUC values stay consistent with using different start times. This shows that the summary features predictive strength is not affected by the start

time.

	0	15	30	45	60	120	180
baseline 1	0.659992	0.661070	0.661265	0.660901	0.660261	0.659855	0.660598
baseline 2	0.811185	0.812167	0.812493	0.812012	0.811513	0.809206	0.806845
summary	0.694657	0.695172	0.694351	0.694907	0.695737	0.691010	0.692128
all	0.837863	0.838777	0.838533	0.838016	0.836148	0.831940	0.828085

Table 5.4: AUC Score in relation to the start of the dialog

# Chapter 6

## Conclusions

This paper explores the use of features that capture speakers' past turn-taking behavior in predicting whether there will be a turn transition. These summary features include (a) relative turn length: how the current turn under construction compares to the current speaker's average turn length; and (b) relative floor control: the percentage of time that the current speaker has held the floor. We include two versions of each, one based on time, and one based on number of words. Relative turn length should capture whether one or both of the speakers tends to hold the turn over multiple utterances, while relative floor control captures whether one speaker is dominating the conversation. Both of these factors should influence who will speak next.

In evaluating our model on data from the Switchboard corpus, we find that our summary features on their own do better than just using the previous speech act (accuracy of 66.14% vs 60.26%). We also find that when we add these features to a model that uses the last two speech acts, we also see an improvement (76.05% vs 74.43%). These results show the potential of modeling speakers' past turn-taking behavior in predicting upcoming turn-transitions. Better modeling of turn-taking should lead to more natural and efficient spoken dialogue systems.

### 6.1 Future Direction

In this work, the local features that we considered in our baseline model were just the last two speech acts. Other work on turn-taking prediction use a richer set of local features, such as syntactic [6, 34, 10, 5, 24, 2], prosodic [6, 10, 40, 8, 5, 31, 30, 16, 2], pragmatic [10, 11, 30], semantic [30] and non-verbal [18]. In future work, it would be good to evaluate the contribution of our summary features with a richer set of local features.

In our work, we evaluated our model on the Switchboard corpus. In future work, it would also be good to evaluate our summary features on other corpora, especially task-based dialogues. Tasks in which there is a difference in the role of the user and speaker, such as in Trains [?], should

benefit from modeling the past turn-taking behavior of each speaker.

In our work, we computed the relative turn length and relative speaker control using the turn length average as computed over all the previous turns. In future work, it would be interesting to use simple moving average (measured over multiple window width) as well as exponential moving average.

More generally, the summary features introduced in this work represent just one aspect of the conversational image of the user. Future work should try to “summarize” other local features by creating the average value of a local feature over past turns. For example, we can compute relative speech rate, or relative use of stereotyped expressions.



# Bibliography

- [1] H. Arsikere, E. Shriberg, and U. Ozertem. Enhanced end-of-turn detection for speech to a personal assistant. In *AAAI Spring Symposium on Turn-taking and Coordination in Human-Machine Interaction*, 2015.
- [2] M. Atterer, T. Baumann, and D. Schlangen. Towards incremental end-of-utterance detection in dialogue systems. In *COLING (Posters)*, pages 11–14, 2008.
- [3] L. Breiman. Random forests. *Machine Learning*, 45(1):5–32, 2001.
- [4] S. Calhoun, J. Carletta, J. M. Brenier, N. Mayo, D. Jurafsky, M. Steedman, and D. Beaver. The nxt-format switchboard corpus: a rich resource for investigating the syntax, semantics, pragmatics and prosody of dialogue. *Language Resources and Evaluation*, 44(4):387–419, 2010.
- [5] J. P. De Ruiter, H. Mitterer, and N. J. Enfield. Projecting the end of a speaker’s turn: A cognitive cornerstone of conversation. *Language*, pages 515–535, 2006.
- [6] S. Duncan. Some signals and rules for taking speaking turns in conversations. *Journal of Personality and Social Psychology*, 23:283–292, 1972.
- [7] L. Ferrer, E. Shriberg, and A. Stolcke. Is the speaker done yet? faster and more accurate end-of-utterance detection using prosody. In *INTERSPEECH*, 2002.
- [8] L. Ferrer, E. Shriberg, and A. Stolcke. A prosody-based approach to end-of-utterance detection that does not require speech recognition. In *ICASSP*, pages 608–611, 2003.
- [9] C. E. Ford. At the intersection of turn and sequence. *Studies in Interactional Linguistics*, 10:51, 2001.
- [10] C. E. Ford and S. A. Thompson. Interactional units in conversation: syntactic, intonational, and pragmatic resources for the management of turns. *Studies in Interactional Sociolinguistics*, 13:134–184, 1996.
- [11] S. Garrod and M. J. Pickering. The use of content and timing to predict turn transitions. *Frontiers in Psychology*, 6, 2015.

- [12] J. J. Godfrey, E. C. Holliman, and J. McDaniel. SWITCHBOARD: Telephone speech corpus for research and development. In *ICASSP*, pages 517–520, 1992.
- [13] J. J. Godfrey, E. C. Holliman, and J. McDaniel. Switchboard: telephone speech corpus for research and development. In *Acoustics, Speech, and Signal Processing, 1992. ICASSP-92., 1992 IEEE International Conference on*, volume 1, pages 517–520 vol.1, Mar 1992.
- [14] A. Gravano and J. Hirschberg. Turn-taking cues in task-oriented dialogue. *Computer Speech & Language*, 25(3):601–634, 2011.
- [15] N. Guntakandla and R. Nielsen. Modelling turn-taking in human conversations. In *AAAI Spring Symposium on Turn-Taking and Coordination in Human-Machine Interaction*, Stanford CA, 2015.
- [16] R. Hariharan, J. Hakkinen, and K. Laurila. Robust end-of-utterance detection for real-time speech recognition applications. In *ICASSP*, pages 249–252, 2001.
- [17] G. Hinton, L. Deng, D. Yu, G. E. Dahl, A.-r. Mohamed, N. Jaitly, A. Senior, V. Vanhoucke, P. Nguyen, T. N. Sainath, et al. Deep neural networks for acoustic modeling in speech recognition: The shared views of four research groups. *IEEE Signal Processing Magazine*, 29(6):82–97, 2012.
- [18] A. Kendon. Some functions of gaze-direction in social interaction. *Acta Psychologica*, 26:22–63, 1967.
- [19] W. J. M. Levelt. *Speaking : from intention to articulation / Willem J.M. Levelt*. MIT Press Cambridge, Mass, 1989.
- [20] S. C. Levinson. Turn-taking in human communication—origins and implications for language processing. *Trends in Cognitive Sciences*, 20(1):6–14, 2016.
- [21] S. C. Levinson and F. Torreira. Timing in turn-taking and its implications for processing models of language. *Frontiers In Psychology*, 6, 2015.
- [22] R. Levitan, S. Benus, A. Gravano, and J. Hirschberg. Entrainment and turn-taking in human-human dialogue. In *AAAI Spring Symposium on Turn-Taking and Coordination in Human-Machine Interaction, Stanford, CA*, 2015.
- [23] L. Magyari, M. C. Bastiaansen, J. P. de Ruiter, and S. C. Levinson. Early anticipation lies behind the speed of response in conversation. *Journal of Cognitive Neuroscience*, 2014.

- [24] L. Magyari and J. P. De Ruiter. Prediction of turn-ends based on anticipation of upcoming words. *Frontiers in Psychology*, 3:376, 2012.
- [25] W. McKinney. pandas: a foundational python library for data analysis and statistics.
- [26] T. Meshorer and P. A. Heeman. Using past speaker behavior to better predict turn transitions. In *INTERSPEECH*, 2016.
- [27] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830, 2011.
- [28] A. Raux. *Flexible turn-taking for spoken dialog systems*. PhD thesis, US National Science Foundation, 2008.
- [29] A. Raux and M. Eskenazi. A finite-state turn-taking model for spoken dialog systems. In *Proceedings of Human Language Technologies: The 2009 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, pages 629–637, 2009.
- [30] A. Raux and M. Eskenazi. Optimizing the turn-taking behavior of task-oriented spoken dialog systems. *ACM Transactions on Speech and Language Processing (TSLP)*, 9(1):1, 2012.
- [31] B. S. Reed. Units of interaction: intonation phrases or turn constructional phrases. *Actes/Proceedings from IDP (Interface Discours & Prosodie)*, pages 351–363, 2009.
- [32] C. Riest, A. B. Jorschick, and J. P. De Ruiter. Anticipation in turn-taking: mechanisms and information sources. *Frontiers in Psychology*, 6, 2015.
- [33] S. G. Roberts, F. Torreira, and S. C. Levinson. The effects of processing and sequence organization on the timing of turn taking: a corpus study. *Frontiers in Psychology*, 6, 2015.
- [34] H. Sacks, E. A. Schegloff, and G. Jefferson. A simplest systematics for the organization of turn-taking for conversation. *language*, pages 696–735, 1974.
- [35] R. Sato, R. Higashinaka, M. Tamoto, M. Nakano, and K. Aikawa. Learning decision trees to determine turntaking by spoken dialogue systems. In *ICSLP*, pages 861–864, 2002.
- [36] E. A. Schegloff and H. Sacks. Opening up closings. *Semiotica*, 8(4), 1973.

- [37] D. Schlangen. From reaction to prediction: Experiments with computational models of turn-taking. *Proceedings of Interspeech 2006, Panel on Prosody of Dialogue Acts and Turn-Taking*, 2006.
- [38] E. O. Selfridge and P. A. Heeman. Importance-driven turn-bidding for spoken dialogue systems. In *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics*, pages 177–185, Uppsala Sweden, 2010.
- [39] M. Selting. The construction of units in conversational talk. *Language in Society*, 29(04):477–517, 2000.
- [40] E. Shriberg, A. Stolcke, D. Hakkani-Tür, and G. Tür. Prosody-based automatic segmentation of speech into sentences and topics. *Speech Communication*, 32(1):127–154, 2000.
- [41] A. Stolcke, N. Coccaro, R. Bates, P. Taylor, C. Van Ess-Dykema, K. Ries, E. Shriberg, D. Jurafsky, R. Martin, and M. Meteer. Dialogue act modeling for automatic tagging and recognition of conversational speech. *Computational linguistics*, 26(3):339–373, 2000.
- [42] M. Tice and T. Henetz. Turn-boundary projection: Looking ahead. In *The 33rd Annual Meeting of the Cognitive Science Society [CogSci 2011]*, pages 838–843. Cognitive Science Society, 2011.
- [43] F. Torreira, S. Bögels, and S. C. Levinson. Breathing for answering: the time course of response planning in conversation. *Frontiers in Psychology*, 6, 2015.
- [44] J. Webber. A programmatic introduction to neo4j. In *Proceedings of the 3rd Annual Conference on Systems, Programming, and Applications: Software for Humanity*, SPLASH ’12, pages 217–218, New York, NY, USA, 2012.
- [45] M. Wilson and T. P. Wilson. An oscillator model of the timing of turn-taking. *Psychonomic Bulletin & Review*, 12(6):957–968, 2005.