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

June 2016

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
Research Assistant Professor

Stephen Wu
Research Associate Professor

NA
NA

Dedication

Acknowledgements

Contents

Dedication . . . . **iv**

Acknowledgements . . . . **v**

Abstract . . . . **vii**

I Introduction **1**

1 Introduction **2**
1.1 Thesis Statement 3
1.2 Approach 3
1.2.1 Local Features 4
1.2.2 Summary Features 4
1.3 Dissertation Structure 5
1.4 Contribution 5

2 Related Work **6**
2.1 Human - Human Conversations 6
2.2 Human - Computer Conversations 8

3 Experiment **10**
3.1 Data 10
3.2 Data Preparation 10
3.3 Results 13

4 Conclusions **15**
4.1 Future Direction 15

Bibliography **17**

Biographical Note **21**

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

June 2016
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. The turn taking system in human conversations supports minimum speaker overlap during turn transition and minimum gaps between turns. Spoken dialogue systems represent new form of conversational user interface, which permit users to use their voice in order to interact with the computer. As such, the turn taking capabilities of SDS should evolve from simple timeout to more human like. Recent advances in turn taking system 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, on average, use long turns, will likely to use long turns in the future. Moreover, speakers with more control of the conversation floor will be unlikely to yield their turn often. As the conversational image of the speaker evolves with the conversation, 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 time and words) relative to the speaker's average turn length. Relative floor control measures the speaker's control of the conversation floor (in time 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 improve turn transitions prediction.

Part I

Introduction

Chapter 1

Introduction

Conversations are the most common form of everyday social interaction. Conversations are interactive and characterized by rapid exchange of messages between the conversants. In their seminal work [32] that created the foundation for the field of conversation analysis, Sack et al. found that turn taking is a universal set of rules which define the framework of interaction. Later work showed that the turn taking system also and crosses culture, age and language. According to [32] the majority of time in a conversation a single speaker control the conversation floor, that the conversant takes turns, and that the gap and overlap between turns is kept to minimum. This attributes apply regardless of turn length (from a single word to an whole sentence) and conversation length. To keep the gaps between turns minimal while supporting speaker change, the listener must prepare its turn before the current turn end and the speaker need to release the conversation floor such that the next speaker could gain control of it. In this study we concern of the latter.

Spoken dialogue systems (SDS) are a computer systems that support conversional user interface. User engage in a conversation with computer in order to perform their needed task (for example information seeking) in using their natural apparatus - voice. Hence to be effective and user friendly, an SDS system should not only be able to understand the semantic of user utterance, but should also adhere to the delicate system of turn exchange and timing.

Up until recently the major barrier to natural interaction with users was high degree of errors during speech recognition. However, recent advances in machine learning , greatly reduced this barrier. This increase the importance of other components in SDS like turn taking subsystem.

Early SDS systems did not contain any turn management component and instead used a fixed timeout to detect user turn ending. Using simple a timeout led to both barge-in situations in which the system prematurely started to speak while the user still assume its turn. In order to decrease the barge-ins an increase in timeout cause large gaps between turns, where the user wait for the system to speak. To improve the situation, later SDS systems are trying to incorporate recent finding in human-human conversation and create prediction models which are based on features from the latest utterance in order to predict turn transition. Prediction in human-human conversation is based syntactic [32, 5], prosodic [10, 7, 8], pragmatic [9].

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

1.1 Thesis Statement

This thesis describe will test the hypothesis that SDS prediction of turn transition can be improved by using summary features in addition to local features. We believe that better prediction will increase the user friendliness of SDS system and lead to more efficient and fluent conversations between human and machines. To test the hypothesis, we choose pragmatics (dialog acts) to represent the user utterance. Our baseline models use current and previous dialog act to predict turn transition. We plan to compute summary features over multiple past turns of the current speaker. The features measure the *relative turn length* of the current turn, and the *relative floor control* of the current speaker. We than train prediction model using the summary feature as well as the local feature and test if the summary feature perform better prediction of turn transition than the baseline model.

1.2 Approach

To test the summary features effectiveness, we used the NXT version of the Switchboard corpus [4, 12] to train random forest models [25]. The baseline models were trained on 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.

This section defines the local and summary features. The local features are based on pragmatics and consist of the current and previous dialog acts. The summary features are based on measurements of each speaker's behavior over over the preceding turns in the dialogue.

1.2.1 Local Features

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

$$\ldots s_{i-2}, d_{i-2}, s_{i-1}, d_{i-1}, s_i, d_i \ldots \tag{1.1}$$

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:

$$\ldots d_{i-2}, y_{i-1}, d_{i-1}, y_i, d_i, y_{i+1} \ldots \tag{1.2}$$

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.

1.2.2 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 denote 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 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

$$avg\_t_i = \frac{\sum_{t \in S_i} length(t)}{|S_i|} \tag{1.3}$$

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).

$$rt_i = \frac{length(t_i)}{avg\_t_i} \tag{1.4}$$

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 again define S_i as above, and we define

L_i to be the turns of the other conversant (the listener of d_i). We first compute the conversation length up to d_i denoted as c_i which excludes inter-utterance pauses.

$$c_i = \sum_{t \in S_i \cup L_i} length(t) \tag{1.5}$$

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:

$$rc_i = \frac{\sum_{t \in S_i} length(t)}{c_i} \tag{1.6}$$

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.

1.3 Dissertation Structure

Chapter 2 present the background information for turn projection in both human-human and human-computer conversations. The chapter describe how current SDS system support turn taking. Chapter 3 describe the local features and the summary features. Chapter 4 describe the experiment which include data preparation and the machine learning method which was used to build the predication models. Chapter 5 show the results which we obtained from the baseline models as well as model which include the summary features. Finally in chapter 4 we discuss the results and present future work.

1.4 Contribution

Our results show that the summary features improve prediction performance in both area under the curve (AUC), 0.65 vs 0.63, and F1, 66.42% vs 54.97%. In addition, the model that was trained on all the features performs better than the local features model in both AUC, 0.82 vs 0.79, and F1, 74.87% vs 74.08%. The results show that using the summary features can help predict turn transitions.

Chapter 2

Related Work

This section present work related to turn transition prediction in both human-human conversation as well human-computer conversations. The importance of turn transition was first established in the seminal work of [32] which form the foundation to the field of conversation analysis. In addition, other early theories recognized the importance of signalling (using both prosodic, syntactic and even non verbal ques) the upcoming release of the conversation floor by the speaker. Since most of the studies in human-human conversations study the affect of features derived from last few utterance on projection of turn transition, the research presented in this disassertion complement both the turn allocation system as presented by [32] and the signaling theory as presented in [**?**].

First generation SDS systems used a timeout based approach to turn taking. This simple approach causes system barge ins (in which the system wrongly identify intra turn pauses as turn transition) as well as large gaps between turns (in which the user finish it turn, but wait until the timeout trigger). Second generation SDS system incorporating more advanced machine learning models which uses various local features - prosodic, syntactic , pragmatics to predict the release of conversation floor. As before, our work complement the local features used by the current machine learning models and suggest that those model can be improved if augmented by summary features.

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 [17] added gaze as a signal to turn transition. Our summary features complement the set of signals as suggested by [6].

Turn allocation was introduced in the seminal work by Sacks, Schegloff, and Jefferson [32], 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. The work on turn bidding is also related [35], which suggested that each conversant measures the importance of their utterance when negotiating the right to the conversation floor.

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 [32], prosodic [10] and pragmatic [10, 9]). Our work augments the local utterance features with summary features that can be used to improve projectablity.

Entrainment was presented in [42], which suggested entrainment of endogenous oscillators in the brains of the speaker and the listener on the basis of the speaker syllabus production. In their study, the speaker and the listener are counter phased such 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 [16] reduce speech recognition error rates, the problem of turn taking in SDS rises in importance. 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 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 last utterance. As different studies use a variety of features, we will outline those that used counting features that are close to the summary features.

Arsikere et al. [2] focus on utterance segmentation in the context of incremental dialog system. Using the switchboard corpus, they used a decision tree algorithm to decide if a word is utterance final using various features and in particular the number of words in the turn so far. The usage of count features improves precision by 10% but has very low recall (7%), which might have occurred, according to the author, from turns with only one word.

Gravano and Hirschberg [13] used the Columbia games corpus in order 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 [28] performed a comprehensive study on features that inform turn changes. The study found that timing features, like 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 [14] used a model based on N grams of dialog acts to predict turn transitions. They trained a decision tree model using the switchboard data and tested bigram, trigram and 4 grams models of dialog acts with and without speaker id. They achieved an F1 measure of 0.67 for the trigram model. In this paper we based our baseline models on bigrams and trigrams of dialog acts. We also mapped the switchboard dialog acts from 148 dialog acts down to 9 in order to reduce data dimensionality. The prediction performance of our baseline model is comparable to their results.

Chapter 3

Experiment

3.1 Data

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

The switchboard corpus is the first large collection of phone conversations who was collected in 1990-1191. The original corpus composed of 2483 calls involving 520 speakers. Each call range from 1.5 minute to 10 minute with an average length of 6 minutes. Conversation involve randomly chosen topic, between two randomly selected speakers. The corpus was later improved and was released as part of the Penn Treebank 3 corpus which included 650 conversations. The current release as was used for this research includes 642 conversations and just over 830000 words.

The current corpus contain different annotations over the original data. At the base we used the token annotation, which include words and pauses between them. The token annotation was used to compute the number of words per turn and the timing of the turn. We used the dialog act annotation to segment the turn. The dialog act annotation was derived from , and include pointer from dialog acts to the syntax tree annotation. We thus derived the dialog act start and stop time from the left most token in the systax tree start time and the right most token stop time.

3.2 Data Preparation

Figure 3.1 shows the experiment data pipeline. Data is imported from the NXT switchboard corpus [4] into a graph database [41]. Figure 3.2 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.

Figure 3.1: The experiment data pipeline

Figure 3.2: Conversation graph data model

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

- We exclude 11 dialogue acts that were coded in Switchboard as “other”.
- Since we believe that it takes a certain amount of time to build a stable conversational image, in evaluating our model, we removed all turns that occurred in the first part of each conversation.

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

Table 3.1: Mapping from dialog act to dialog act class

For this paper, we used an estimate of 120 seconds. This reduced the number of dialog acts from 50,633 to 37,508.

- 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}.

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, it 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.

To find the optimal hyper parameters, we ran a grid search over the *max_features* and *max_depth* hyper parameters for each model. The hyper parameters search was done over $\{sqrt, log2, 10\}$ for *max_features* and $\{5, 7, 9\}$ for *max_depth*. When training the model, we used the optimal hyper parameters for each feature set.

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.

3.3 Results

We first analyze the results in terms of accuracy: how often the models correctly predicted whether a turn transition occurred or not, in other words, how often it predicts the correct value of y_{i+1}. Table 2 shows the results of training a random forest for each model. We see that using the summary features provides better accuracy than baseline 1, which only uses the current current dialog act (66.14% vs 60.26%). In addition, using the full model yields an improvement of over 1.58% in the result.

Model	Accuracy	AUC	hyper parameters
Baseline 1	60.26%	0.63	max_features=sqrt, max_depth=7
Baseline 2	74.43%	0.79	max_features=log2, max_depth=9
Summary	66.14%	0.65	max_features=sqrt, max_depth=5
Full	76.05%	0.82	max_features=10, max_depth=9

Table 3.2: Accuracy and AUC results

The effect can also be seen in Figure 3, which shows the ROC curves and the AUC for each model. We notice that the AUC of the summary model is better than baseline model (0.65 vs 0.63), and when adding the summary features to the local features, the full model, we see the AUC improves (0.82 vs 0.79). This suggests that while the discrimination facility of the summary features is lacking (AUC < 0.7), adding them to a classifier that uses local features (full model) yields better results than the baseline.

In addition to analyzing the results in terms of accuracy, we also analyze the results of the four models in terms of how well we predict that there is a change in speaker (i.e., y_{i+1} indicates that there was a turn switch). Table 3 shows the results in terms of recall, precision, and F1, which combines the two scores. Although 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.

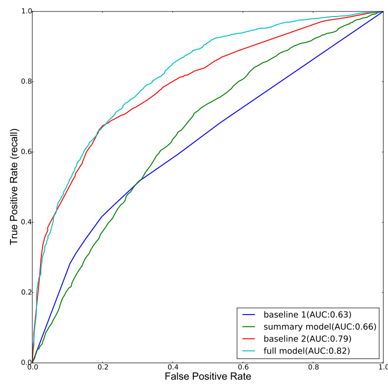

Figure 3.3: ROC curves and AUC of the different models

Model	Precision	Recall	F1
Baseline 1	69.49%	45.52%	54.97%
Baseline 2	80.38%	68.80%	74.08%
Summary	64.55%	68.88%	66.42%
Full	76.17%	77.25%	74.87%

Table 3.3: Precision, recall and F1 results

Chapter 4

Conclusions

This paper explores the use of features that capture speakers' past turn-taking behavior in predicting whether their 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.

4.1 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, 32, 10, 5, 23, 2], prosodic [6, 10, 37, 8, 5, 29, 28, 15, 2], pragmatic [10, 11, 28], semantic [28] and non verbal [17]. 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.

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

[14] 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.

[15] R. Hariharan, J. Hakkinen, and K. Laurila. Robust end-of-utterance detection for real-time speech recognition applications. In *ICASSP*, pages 249–252, 2001.

[16] 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.

[17] A. Kendon. Some functions of gaze-direction in social interaction. *Acta Psychologica*, 26:22–63, 1967.

[18] W. J. M. Levelt. *Speaking : from intention to articulation / Willem J.M. Levelt.* MIT Press Cambridge, Mass, 1989.

[19] S. C. Levinson. Turn-taking in human communication–origins and implications for language processing. *Trends in Cognitive Sciences*, 20(1):6–14, 2016.

[20] S. C. Levinson and F. Torreira. Timing in turn-taking and its implications for processing models of language. *Frontiers In Psychology*, 6, 2015.

[21] 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.

[22] 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.

[23] L. Magyari and J. P. De Ruiter. Prediction of turn-ends based on anticipation of upcoming words. *Frontiers in Psychology*, 3:376, 2012.

[24] W. McKinney. pandas: a foundational python library for data analysis and statistics.

[25] 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.

[26] A. Raux. *Flexible turn-taking for spoken dialog systems.* PhD thesis, US National Science Foundation, 2008.

[27] 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.

[28] 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.

[29] B. S. Reed. Units of interaction: intonation phrases or turn constructional phrases. *Actes/Proceedings from IDP (Interface Discours & Prosodie)*, pages 351–363, 2009.

[30] C. Riest, A. B. Jorschick, and J. P. De Ruiter. Anticipation in turn-taking: mechanisms and information sources. *Frontiers in Psychology*, 6, 2015.

[31] 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.

[32] H. Sacks, E. A. Schegloff, and G. Jefferson. A simplest systematics for the organization of turn-taking for conversation. *language*, pages 696–735, 1974.

[33] 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.

[34] 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.

[35] 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.

[36] M. Selting. The construction of units in conversational talk. *Language in Society*, 29(04):477–517, 2000.

[37] 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.

[38] 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.

[39] 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.

[40] F. Torreira, S. Bögels, and S. C. Levinson. Breathing for answering: the time course of response planning in conversation. *Frontiers in Psychology*, 6, 2015.

[41] 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.

[42] M. Wilson and T. P. Wilson. An oscillator model of the timing of turn-taking. *Psychonomic Bulletin & Review*, 12(6):957–968, 2005.

Biographical Note

Morbi luctus, wisi viverra faucibus pretium, nibh est placerat odio, nec commodo wisi enim eget quam. Quisque libero justo, consectetuer a, feugiat vitae, porttitor eu, libero. Suspendisse sed mauris vitae elit sollicitudin malesuada. Maecenas ultricies eros sit amet ante. Ut venenatis velit. Maecenas sed mi eget dui varius euismod. Phasellus aliquet volutpat odio. Vestibulum ante ipsum primis in faucibus orci luctus et ultrices posuere cubilia Curae; Pellentesque sit amet pede ac sem eleifend consectetuer. Nullam elementum, urna vel imperdiet sodales, elit ipsum pharetra ligula, ac pretium ante justo a nulla. Curabitur tristique arcu eu metus. Vestibulum lectus. Proin mauris. Proin eu nunc eu urna hendrerit faucibus. Aliquam auctor, pede consequat laoreet varius, eros tellus scelerisque quam, pellentesque hendrerit ipsum dolor sed augue. Nulla nec lacus.