Investigating Linguistic and Semantic Features for Turn-Taking Prediction in Open-Domain Human-Computer Conversation

S. Zahra Razavi[1] *, Benjamin Kane*[2]*, Lenhart K. Schubert*[3]

[1,2,3]University of Rochester

srazavi@cs.rochester.edu, bkane2@u.rochester.edu, schubert@cs.rochester.edu

Abstract

In this paper we address the problem of turn-taking prediction in open-ended communication between humans and dialogue agents. In a non-task-oriented interaction with dialogue agents, user inputs are apt to be grammatically and lexically diverse, and at times quite lengthy, with many pauses; all of this makes it harder for the system to decide when to jump in. As a result recent turn-taking predictors designed for specific tasks or for human-human interactions will scarcely be applicable. In this paper we focus primarily on the predictive potential of linguistic features, including lexical, syntactic and semantic features, as well as timing features, whereas past work has typically placed more emphasis on prosodic features, sometimes supplemented with non-verbal behaviors such as gaze and head movements. The basis for our study is a corpus of 15 "friendly" dialogues between humans and a (Wizard-of-Oz enabled) virtual dialogue agent, annotated for pause times and types. The model of turn-taking obtained by supervised learning predicts turn-taking points with increasing accuracy using only prosodic features, only timing and speech rate features, only lexical and syntactic features, and achieves state-of-the art performance with a mixture-of-experts model combining these features along with a semantic criterion.

Index Terms: turn-taking, human-computer conversation

1. Introduction

Spoken dialogue systems have been getting more common in everyday life applications ranging from mobile assistants and customer service to conversational companies designed to chat with users on a variety of topics. However, smooth turn exchange behavior in human-machine interaction is still an issue for such systems. Using a silence threshold, which is the most common turn-taking strategy, can easily result in long awkward silences or frequent overlaps and confusion. The situation can be worse in open-domain dialogue systems where users might provide arbitrarily long inputs with many pauses without willing to yield the turn.

Turn-taking behavior has been an area of research since early 1970 [1] by studying cues people use in their interaction including lexical, prosodic, and gestural. Later researchers tried to propose predictive models so that machines could efficiently decide users end-of-turns. These models showed some success in domain-specific tasks such as games [2] and map task [3]. Other researchers focused on predicting exchange points in human-human interactions such as Switchboard [4, 5] by tracking cues from both sides of interaction. In the past couple years, some started applying sequential models such as LSTM with focus on prosodic features to develop predictive models [6, 7, 8].

In this paper, we address the problem of turn-taking in open-domain conversation between a dialogue agent and users using a variety of users speech and language features. Instead of relying on human-human interactions, we collected data from actual casual dialogue between users and machine. The silences were annotated and five categories of features were extracted including prosodic, timing, lexical, syntactic and semantic. We study the impact of each type on the prediction power and discussed the results. Then we propose a two-layer context-aware combination model which can learn to efficiently combine different aspects of users input. The model performance is comparable with the state of the art systems.

2. Literature Review

The ability to handle smooth turn exchanges in human-human conversation is universal among all language speakers, with gaps between speaker turns being held to around 200ms on average [9]. Previous experiments suggest that people listening to non-faulty speech generally predict end-of-turns about 1200ms before they happen [10]. Speakers provide many types of prosodic, linguistic, and nonverbal cues that enable such turn predictions, most notably a higher pitch slope and higher mean pitch and intensity before a turn-ending [11, 2]; longer average syllable length [12], gaze direction [13], and certain types of vocabulary to indicate turn-holding or turn-yielding [14]; and it is also thought that syntactic and semantic completion can provide clues for detecting end-of-turns, suggesting features such as POS tags, size and type of the last phrase in a turn, etc. [2].

Compared with human-human conversation, smoothly automating human-computer conversation poses some unique challenges. Some studies have found that people tend to adapt their behavior when a machine is slower than a person in responding to end-of-turns [11, 15]. Studies of human-computer turn-taking have generally sought to develop automated methods of predicting turns based on conversation cues. In one analysis of a map-task dialogue system, a combination of prosodic and contextual features was found to predict turn-taking points with 66% accuracy, and lexico-syntactic features were found to predict such points with 84% accuracy [3]. However, this study concerned a task-oriented system, as opposed to open-ended dialogue, so the vocabulary and types of exchanges were much narrower in scope and turn-taking less variable than in our system. Conversation histories have also been found to be an effective feature, with one study on Switchboard human-human dialogue predicting turn-taking using a Random Forest with features such as previous turn length and floor control, and achieving an F1 of 74% [5].

More recent studies have attempted to use recurrent neural networks with combinations of acoustic/prosodic features and some linguistic features such as word embeddings and POS tags [6, 16, 17]. A high F1 score of 85.5% was achieved in a study on corpora of human-human interaction by using a multiscale LSTM with acoustic and linguistic features, and including gaze features improved this score to 93.5% [18]. Similar studies us-

ing LSTM classifiers have been done using Japanese corpora as well [19, 20, 8, 21]. These studies focused primarily on prosodic features, although word embeddings were also used in some.

Turn-taking has also been found to depend on the specific task (e.g., transportation planning vs. topical chatting) and speech act of a user [22], and one study attempted to extend an LSTM model to predict a speaker's intentions along with turn-taking [4]. However, it's not clear if these results would hold in an open-ended human-machine interaction as opposed to task-oriented dialogue. Another study also attempted to predict backchannels and fillers as well as turn-taking using prosody [7]. A general observation about prior studies is that F-scores for turn prediction depend very much on the scope of the dialogues (e.g., map task: 81.7 [17] vs. Switchboard: 65.8 [4]), the size of the training corpus (e.g., 2.5 hours, job interviews: 77.3 [7] vs. 11 hours, MAHNOB: 93.4 [18]), as well as what is being measured and predicted (e.g., use of visual as well as linguistic features, or inclusion/ exclusion of backchannels as turns). Also as pointed out in [22], in more difficult tasks pauses may be due to thinking about what to say rather than whether to yield the turn. We are not aware of any studies of turn taking in topically broad human-computer dialogues based on the Wizard-of-Oz (WOZ) technique, other than perhaps the "robotic" job interviews [7, 8] just cited.

3. Data preparation

We used a corpus of 15 subjects interacting with the LISSA conversational agent collected in a previous WOZ study [23]. The users were all native English speakers between the ages of 18 and 25. During the conversation the virtual agent leads casual conversation on different topics such as "getting to know each other", "hobbies", "movies", "food", etc. Each conversation contains 15-25 turns on each side. There are some interruptions and moments of speech overlap but most turn exchanges in the data happen smoothly. After collecting the transcripts, we marked the silences longer than 500 milliseconds using Praat [24]. Following the convention in the field, we call an utterance between two pauses an "Inter Pausal Unit" or IPU. We obtained 1099 silence points, and asked three undergraduate RAs to annotate these points in the transcripts with the occurrence time and duration. RAs also labeled the silences with four categories, based just on the transcripts:

- "Turn-holding" (TH) means that the user is not semantically or syntactically done with what they are saying so that it would be inappropriate for the avatar to try to take the turn at that silence point.

- "Potential end-of-turn" (PET) means the silence point can be regarded as a turn-taking point by the avatar, although it was not an actual turn exchange point in the conversation. In other words, there is no semantic or syntactic incompleteness, and it would not be particularly inappropriate if the virtual agent tried to take the turn at such times.

- "End-of-turn" (ET) means that it was an actual end-of-turn point in the conversation and it was a smooth one, so there was no interruption or overlapping speech.

- "Interruption" (INT) means that the avatar interrupted the user without letting the user finish. So, INT points are turn-exchange points but not smooth ones.

Based on the above labeling we ended up with 537 strong turn-holding (TH), 267 end-of-turn (ET), and 263 potential end-of-turn (PET) points. The Fleiss kappa score for the subjective labels, "TH" and "PET", was 0.86, indicating substantial inter-annotator agreement. As the PET points were judged to be appropriate points for machine to take the turn, we count them as end-of-turn for prediction purposes. Also we remove the interruption points in order to deal with smooth exchange data. As a result the problem turns into two class prediction.

Table 1: *Data statistics– TH: turn-holding, PET: potential end-of-turn, ET: end-of-turn, and INT: interruption*

General data statistics	Number of dialogues	15
	Number of users' turns	301
	Number of silence points	1099
Pause types statistics	Number of TH silences	537
	Number of PET silences	263
	Number of ET silences	267
	Number of INT points	33
Users' turns statistics	Min length (sec)	0.26
	Max length (sec)	75.2
	Length average (sec)	13.32

The data shows that the users' turn lengths varied between 0.26 seconds (one word) to 75.23 seconds (232 words) with an average of 13.32 seconds (sd = 12.04), where we observed up to 12 pauses in a user turn. Also, we observed that 31 percent of the strong turn-holding points last longer than 1 second where the longest one was 4.95 seconds. These all prove the need for an effective turn predictor while at the same time demonstrating the difficulty of obtaining one. Table 1 shows a summary of the data we collected.

4. Experimental evaluation

The open-domain human-machine data described in the previous section was used for training a model of turn-taking prediction. In order to come up with a model for a turn-taking predictor we collected tens of features associated with silence intervals. We group these features into five categories: prosodic, timing, lexical, syntactic, and semantic features. Some features from the first four categories have been studied in the literature concerned with predicting users' end-of-turn points in human-machine conversation. However, while we know that semantic features play a role in human-human turn-taking behavior [25], they have played no role in end-of-turn predictors, except for some use of domain-specific semantic word tags [3].

In this section we first test the predictive power of different feature categories to gain some insight into the most effective ones. For this we explore Gaussian Nave Bayes (NB) as a generative model and two discriminative models: CART decision tree classifier [26] and Support Vector Machine (SVM, with radial basis kernel function). We compare the performances of these models against the majority class baseline obtained by the ZeroR classifier which was 48.5% correct on average. Then we try two combination models using the most effective features of all categories. For all these classifiers we have used the implementations available in the scikit toolkit [27]. All results presented here are determined using 10-fold cross-validation.

4.1. Individual feature categories

4.1.1. Prosodic Features

Prosodic features are the most commonly used clues for turn exchange prediction. Although they don't show high performance on their own when used in machine learning models, recent efforts to use them in sequence models have led to better

performance, as we discussed in section 2. In this paper we take account of intensity and pitch features, based on previous evidence for their effectiveness. We measure both slope and mean in three different intervals including: last 200ms, last 500ms, and the entire IPU before the pause points.

We sampled pitch and intensity at 10ms using Praat [24], then z-normalized the value for each user. The mean and slope of pitch and intensity over the last 200ms, 500ms and over the whole IPU preceding each silence point were calculated and added as features. Table 2 illustrates the individual and collective performance of various prosodic features.

Table 2: *Accuracy of EOT prediction using prosodic features*

Features	**SVM**	**DT**	**NB**
Pitch mean	52.51	50.92	50.61
Pitch slope	52.13	50.94	50.15
Pitch mean + pitch slope	53.45	51.51	50.51
Intensity mean	54.96	51.95	54.01
Intensity slope	52.15	51.89	55.11
Intensity mean + intensity slope	54.18	51.61	**56.11**
Pitch mean + int. mean + int. slope	**58.15**	51.72	52.71

The results indicate that pitch and intensity mean and intensity slope contribute to predictive power. This is in line with previous studies in task-oriented interactions [3]; however, predictive power seems to be diminished by the more diverse speech patterns people use in open-domain dialogues.

4.1.2. Timing and speaking rate

For each silence point, turn and IPU length features were extracted along with their mean values over the dialogue history for the individual user. Moreover, the average speaking rate in the preceding IPU and its average value for the user were measured in syllables per second. The accuracy values of the prediction model are reported in table 3.

Table 3: *Accuracy of EOT prediction using timing and speaking rate features*

Features	**SVM**	**DT**	**NB**
Turn length	54.79	49.51	51.79
Turn length ratio	56.11	55.23	55.41
IPU length	55.51	51.93	50.59
IPU length ratio	53.01	51.22	50.43
Speaking rate	**62.9**	58.47	60.95
Speaking rate ratio	62.01	59.43	62.26
Turn len. ratio, IPU len., speaking rate	**63.67**	58.52	60.49

The results suggest that turn length ratio has a bigger impact than the absolute turn-length value. This means that turn length should be seen as a user-specific feature rather than a global feature. In other words we should take into account the previous user's behavior in terms of verbosity, for meaningful use of turn length for EOT prediction. Moreover, speaking rate is another impactful feature based on the results. [2] previously showed that speaking rate tends to increase towards turn boundaries in a game task dialogue.

4.1.3. Lexical features

Because of the diverse vocabulary used in open-domain conversations, using content words as features was not feasible. Instead, we relied on some relevant word categories. Here we show the result of using two classes including filler words (filled pauses) and discourse markers. We checked for their appearance right before the pause point.

Table 4: *Accuracy of EOT prediction using lexical features*

Features	**SVM**	**DT**	**NB**
Filler words	57.73	57.73	57.73
Filled pauses	61.69	61.69	61.69
Filler words, filled pauses	**69.73**	**69.73**	**69.73**

It can be seen that both filler words and filled pauses offer significant contributions to a predictor. A closer look into the data shows that people use filled pauses as a very clear signal indicating their desire to hold the turn. Some filler words such as "well", "so", "but", "or", etc., serve the same function.

4.1.4. Syntactic features

Syntactic features have shown strong predictive power for turn-taking in specific-task domains [3]. Here we collect the part-of-speech (POS) tags of two words before the pause point. For this we used the NLTK toolkit, where we mapped the 36 Treebank tags to the reduced set of 17 universal POS tags [28].

Table 5: *Accuracy of EOT prediction using syntactic features*

Features	**SVM**	**DT**	**NB**
Last word POS tag	**68.26**	68.24	63.1
Last two words POS tags	67.22	67.78	62.97

A closer look into the data shows that the most frequent part-of-speech of the words preceding an end-of-turn are Noun, Adjective, Verb, Adverb, Personal pronoun, while the most frequent ones preceding a turn-holding point are Noun, Verb, Coordinating conjunction, Adverb, Preposition.

4.1.5. Semantic Features

According to some studies (e.g., [25]) semantic features can contribute to turn-taking prediction, especially as they might be more robust to poor acoustic conditions. However, there is no simple way to formalize a semantic analysis of conversation [29]. In this paper we study the role of semantic completion. In the type of casual dialogue exemplified by our data, this can be seen as observing some anticipated response to the question asked by the virtual agent. However, automatically recognizing completion of such a response is challenging since it depends on understanding users' inputs in the context of the ongoing dialogue. Here, we capture the semantic content of users' inputs using the dialogue manager designed to automatically lead meaningful conversation with users [30]. At each pause, the dialogue manager extracts one or more "gist-clauses" which are simple explicit English version of users' inputs. These are extracted using context dependent pattern transduction trees. As features, we collect the number of extracted gist-clauses at each silence, the time since last gist-clause was extracted, and a binary feature showing if the last extracted gist was a question.

The results of using these semantic features in a prediction model can be seen in Table 6. It is worth noting that the feature was not available for almost a third of the data points due to the limitation of dialogue manager at the time of data collection.

Table 6: *Accuracy of EOT prediction using semantic features*

Features	SVM	DT	NB
Num. of extracted gist-clauses	58.12	58.15	55.01
Time since last gist-clause	53.71	54.93	49.59
Num. of extracted gist-clauses, Time since last gist-clause	**58.35**	57.46	56.32

Yet, we observe a significant contribution only by leveraging the semantic features. Moreover, a closer look into the predictor output indicates the existence of meaningful pattern extracted by the model, for instance that the system did not consider taking the turn before seeing extraction of any gist-clause.

4.2. Combined model

The results reported above provided initial insight into the various aspects of speech and language that are influential in predicting end-of-turn points. We then worked on designing a two-layer combined model to improve the prediction accuracy.

4.2.1. Simple combined model

The simple combined model consists of the 12 most powerful features of all categories –listed in table 7, based on the results of section 4.1 implemented using three algorithms: SVM, Decision tree, and Gaussian Naive Bayes. The best classifier was SVM with 73.2% accuracy while the Decision tree and Gaussian Naive Bayes achieved 68.71% and 69.54% respectively.

Table 7: *Features used in the combined models*

Feature category	Features
Prosodic	pitch mean, intensity mean, intensity slope
Timing	turn length ratio, IPU length, speaking rate
Lexical & syntactic	filler words, filled pauses, last POS
Semantic	no. of gists, time since last gist, question gist

4.2.2. Two-layer classifier model

We designed a combined model of *k* classifiers, each trained on a subset of the feature space. The idea is inspired by the Mixture of Experts algorithm [31], but instead of different subsets of data, each classifier sees a subset of features. A high-level architecture of the model is shown in Figures 1 with k classifiers, each intended as an expert on a subset of features and a gate responsible for deciding which classifier should be trusted more for any input feature. The gate unit is trained to learn the best way to combine the expert's decisions having the input features. In general, for a mixture of experts model, we have:

$$P(y \mid x, \Theta) = \sum_{i=1}^{k} P(i|x, \Theta_g) P(y|i, x, \Theta_e) \qquad (1)$$

where x is the input, y is the output, *k* is the number of experts, and Θ denotes the parameters of the model, consisting of Θ_g, the parameters of the gate unit, and Θ_e, the parameters of classifiers. In our case, we train each classifier $d_i(\vec{x}^i)$ on a reduced dimensional version of input, $\vec{x}^i$. The final decision is made by combining classifier decisions based on the weights coming from the gate:

$$P(y = \hat{y}) = \sum_{i=1}^{k} g_i(\vec{x}) d_i(\vec{x}^i) = \vec{g}(\vec{x}).\vec{d}(\vec{x}) \qquad (2)$$

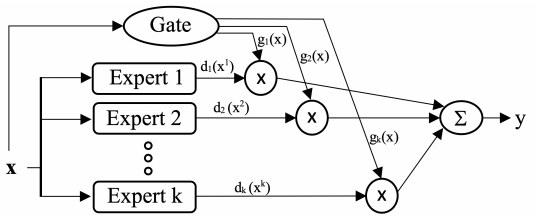

Figure 1: *High-level schema for a mixture-of-experts classifier*

where for input $\vec{x}_{m\times 1}$, the gate outputs $\vec{g}_{k\times 1}(\vec{x}) = (g_1(\vec{x}), g_2(\vec{x}), ..., g_k(\vec{x}))^T$, while the augmented output of classifiers is $\vec{d}_{k\times 1}(\vec{x}) = (d_1(\vec{x}^1), d_2(\vec{x}^2), ..., d_k(\vec{x}^k))^T$. This is interpretable as a weighted vote of all decisions. For the gate module, various structures have been proposed in the literature; here we pick linear structure: $\vec{g}_{k\times 1}(\vec{x}) = \Lambda_{k\times m}.\vec{x}_{m\times 1}$.

For turn-taking prediction, we use the five experts mentioned in section 4.1; by merging lexical and syntactic classifiers, we end up with four base classifiers presented in table 7. The experts were trained on the corresponding reduced dimensions of half of the training data. The second half of the training data were used for learning the combination layer parameters, Λ, using linear regression. The accuracy and F-score of the two-layer combined model is compared with the simple global classifier in table 8.

Table 8: *Accuracy of EOT prediction using combined models*

Features	Accuracy	F1 score
Simple combined model	73.2	72.92
Mixture of experts	**76.47**	**75.72**

5. Discussion and Conclusion

We introduced a data-driven approach for end-of-turn detection using data from open-domain human-machine conversations. We evaluated the respective contributions of prosodic, timing, lexical, syntactic, and semantic features to a predictive model, and found lexical and syntactic features to be the most powerful turn-taking predictors. We also introduced semantic completion as a strong predictor of turn-holding points. We suggested a two-layer context-aware model inspired by mixture-of-experts method to combine the predictors trained on different feature categories of the data. The two-layer structure enhanced the performance compared to simply combining all impactful features. The accuracy and F1-score of the combined model is comparable with some recent attempts on similar tasks such as the "ERICA" WOZ job interviews [7], which also used a relatively small corpus, and a little better than some recent large-corpora studies using Switchboard data [5, 4]. Although such comparisons are of limited significance because of the many factors (discussed in 2) that affect turn-taking behavior and prediction, these results are encouraging given the open-endedness and complexity of our dialogue setting.

6. Acknowledgements

This work was supported by DARPA CwC subcontract W911NF-15-1-0542, and benefited from help by Shuwen Zhang.

7. References

[1] H. Sacks, E. A. Schegloff, and G. Jefferson, "A simplest systematics for the organization of turn taking for conversation," in *Studies in the organization of conversational interaction*. Elsevier, 1978, pp. 7–55.

[2] A. Gravano and J. Hirschberg, "Turn-taking cues in task-oriented dialogue," *Computer Speech Language*, vol. 9, no. 3, pp. 601–634, 2011.

[3] R. Meena, G. Skantze, and J. Gustafson, "Data-driven models for timing feedback responses in a map task dialogue system," *Computer Speech Language*, vol. 28, no. 4, pp. 903–922, 2014.

[4] Z. Aldeneh, D. Dimitriadis, and E. M. Provost, "Improving end-of-turn detection in spoken dialogues by detecting speaker intentions as a secondary task," *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 6159–6163, 2018.

[5] T. Meshorer and P. A. Heeman, "Using past speaker behavior to better predict turn transitions," in *Interspeech*, 2016, pp. 2900–2904.

[6] G. Skantze, "Towards a general, continuous model of turn-taking in spoken dialogue using lstm recurrent neural networks," in *18th Annual Meeting of the Special Interest Group on Discourse and Dialogue (SigDial)*, 2017, pp. 220–230.

[7] K. Hara, K. Inoue, K. Takanashi, and T. Kawahara, "Prediction of turn-taking using multitask learning with prediction of backchannels and fillers," in *Interspeech*, 2018.

[8] D. Lala, K. Inoue, and T. Kawahara, "Evaluation of real-time deep learning turn-taking models for multiple dialogue scenarios," in *20th ACM International Conference on Multimodal Interaction (ICMI)*, 2018, pp. 78–86.

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

[10] H. Wesselmeier, S. Jansen, and H. M. Mller, "Influences of semantic and syntactic incongruence on readiness potential in turn-end anticipation," *Frontiers in Human Neuroscience*, vol. 8, 2014.

[11] A. Raux and M. Eskenazi, "Optimizing the turn-taking behavior of task-oriented spoken dialog systems," *ACM Trans. on Speech and Language Processing (TSLP)*, vol. 9, no. 1, pp. 1–23, 2012.

[12] J. Local and G. Walker, "How phonetic features project more talk," *Journal of the International Phonetic Association*, vol. 42, no. 3, pp. 255–280, 2012.

[13] S. Bgels and S. C. Levinson, "The brain behind the response: Insights into turn-taking in conversation from neuroimaging," *Research on Language and Social Interaction*, vol. 50, no. 1, pp. 71–89, 2017.

[14] F. Donnarumma, H. Dindo, P. Iodice, and G. Pezzulo, "You cannot speak and listen at the same time: A probabilistic model of turn-taking," *Biological Cybernetics*, vol. 111, no. 2, pp. 165–183, 2017.

[15] D. DeVault, J. Mell, and J. Gratch, "Toward natural turn-taking in a virtual human negotiation agent," in *AAAI Spring Symposium on Turn-taking and Coordination in Human-Machine Interaction*, 2015.

[16] A. Maier, J. Hough, and D. Schlangen, "Towards deep end-of-turn prediction for situated spoken dialogue systems," in *Interspeech*, 2017, pp. 1676–1680.

[17] M. Roddy, G. Skantze, and N. Harte, "Investigating speech features for continuous turn-taking prediction using lstms," *Proc. Interspeech 2018*, pp. 586–590, 2018.

[18] ——, "Multimodal continuous turn-taking prediction using multiscale rnns," in *International Conference on Multimodal Interaction (ICMI)*, 2018, pp. 186–190.

[19] R. Masumura, T. Asami, H. Masataki, R. Ishii, and R. Higashinaka, "Online end-of-turn detection from speech based on stacked time-asynchronous sequential networks," in *Interspeech*, 2017.

[20] R. Masumura, T. Tanaka, A. Ando, R. Ishii, R. Higashinaka, and Y. Aono, "Neural dialogue context online end-of-turn detection," in *19th Annual Meeting of the Special Interest Group on Discourse and Dialogue (SigDial)*, 2018, pp. 224–228.

[21] C. Liu, C. T. Ishi, and H. Ishiguro, "Turn-taking estimation model based on joint embedding of lexical and prosodic contents," in *Interspeech*, 2017, pp. 1686–1690.

[22] P. A. Heeman and R. Lunsford, "Turn-taking offsets and dialogue context," in *Interspeech*, 2017, pp. 1671–1675.

[23] M. R. Ali, D. Crasta, L. Jin, A. Baretto, J. Pachter, R. D. Rogge, and M. E. Hoque, "Lissa–Live Interactive Social Skill Assistance," in *2015 International Conference on Affective Computing and Intelligent Interaction (ACII)*. IEEE, 2015, pp. 173–179.

[24] P. Boersma *et al.*, "Praat, a system for doing phonetics by computer," *Glot international*, vol. 5, 2002.

[25] A. Raux and M. Eskenazi, "Optimizing endpointing thresholds using dialogue features in a spoken dialogue system," in *Proceedings of the 9th SIGdial Workshop on Discourse and Dialogue*. Association for Computational Linguistics, 2008, pp. 1–10.

[26] D. Steinberg and P. Colla, "Cart: classification and regression trees," *The top ten algorithms in data mining*, vol. 9, p. 179, 2009.

[27] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg *et al.*, "Scikit-learn: Machine learning in python," *Journal of machine learning research*, vol. 12, no. Oct, pp. 2825–2830, 2011.

[28] S. Petrov, D. Das, and R. McDonald, "A universal part-of-speech tagset," *arXiv preprint arXiv:1104.2086*, 2011.

[29] B. Oreström, *Turn-taking in English conversation*. Krieger Pub Co, 1983, vol. 66.

[30] S. Z. Razavi, L. K. Schubert, M. R. Ali, and M. E. Hoque, "Managing casual spoken dialogue using flexible schemas, pattern transduction trees, and gist clauses," in *Proceedings of the Fifth Annual Conference on Advances in Cognitive Systems (ACS)*, 2017.

[31] S. E. Yuksel, J. N. Wilson, and P. D. Gader, "Twenty years of mixture of experts," *IEEE transactions on neural networks and learning systems*, vol. 23, no. 8, pp. 1177–1193, 2012.