INTERSPEECH 2019
September 15–19, 2019, Graz, Austria

Turn-taking Prediction Based on Detection of Transition Relevance Place

Kohei Hara, Koji Inoue, Katsuya Takanashi, and Tatsuya Kawahara

School of Informatics, Kyoto University, Kyoto, Japan
{hara, inoue, takanashi, kawahara}@sap.ist.i.kyoto-u.ac.jp

Abstract

We address turn-taking prediction in which spoken dialogue systems predict when to take the conversational floor. In natural conversations, many turn-taking decisions are arbitrary and subjective. In this study, we propose taking into account the concept of the transition relevance place (TRP) for turn-taking prediction. TRP is defined as a timing when the current speaking turn can be completed and other participants are able to take the turn. We conducted annotation of TRP on a human-robot dialogue corpus, ensuring the objectivity of this annotation among annotators. The proposed turn-taking prediction model adopts a two-step approach that detects TRP at first and then predicts a turn-taking event if TRP is detected. Experimental evaluations demonstrate that the proposed model improves the accuracy of turn-taking prediction by incorporating TRP detection.

Index Terms: turn-taking, transition relevance place, neural networks, spoken dialogue systems

1. Introduction

Smooth turn-taking is a challenging issue for spoken dialogue systems in order to converse with users in the manner of natural dialogue. A majority of spoken dialogue systems deployed in commercial devices such as smart speakers and smartphone apps assume a fixed-length silence at the end of user speech to trigger the system response. This protocol is applicable and practical when a user utterance is a short query. On the other hand, when we design natural human-robot and human-agent dialogue systems, in which users can speak many utterances per one turn [1, 2, 3, 4, 5, 6], it is necessary to properly predict if the system will take the floor or not at the end of current user speech. Other studies suggested that the smooth and precise turn-taking contributes to increasing users' engagement and satisfaction on dialogue [7, 8].

A number of studies on turn-taking prediction has been done to investigate various feature sets and prediction models. In general, turn-taking prediction is done at the end of each utterance and is formulated as a binary classification: turn-switch or turn-holding. The majority of investigated feature sets are based on prosodic features such as fundamental frequency (F0) and power [9, 10, 11, 12, 13]. Linguistic features were also investigated such as syntactic structure, turn-ending markers, and language model [14, 15]. Moreover, multi-modal features were also considered such as eye-gaze [16, 17, 18, 19], respiration [20, 21, 22], and head-direction [16, 23]. The prediction model was based on conditional random field [16], support vector machines [24], and neural networks [25]. A recent approach is to use recurrent neural networks such as long short-term memory (LSTM), which can handle long-range context of the input sequence, and it achieved higher accuracy than conventional methods [15, 26, 19, 27, 28, 29, 30]. However, the performance is still low in natural conversations. One of the major problems is, even in human conversations, many turn-taking decisions are arbitrary and subjective. For example, when a speech segment is given, it is possible that some persons take a turn and others do not [24].

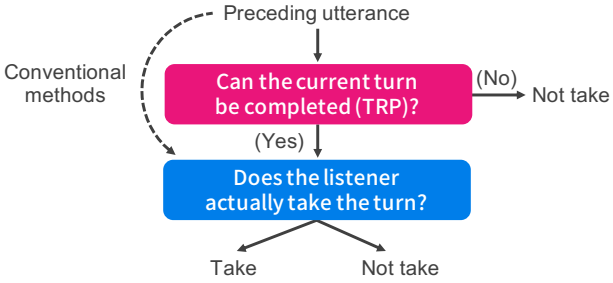

Figure 1: *Relationship between TRP and actual turn-taking event*

In this study, we propose taking into account objective criterion, specifically, the concept of transition relevance place (TRP). TRP is defined as a timing when the current turn can be completed and other participants are able to take the turn [31]. Earlier studies investigated the relationship between TRP and turn-taking cues such as prosodic and syntactic information [32, 33]. We assume that turn-taking decision can be decomposed into two-step decisions as illustrated in Figure 1. At first, a listener judges if the current turn can be completed (TRP). If it is true, then the listener decides to actually take the turn. While conventional methods directly predicted the turn-taking event, our approach is to distinguish turn-holding events (*Not take*) in *TRP* from those in *not TRP* because the required judgment would be different in each situation. Based on this assumption, we propose a two-step turn-taking prediction: TRP detection and turn-taking prediction at TRP. By separately designing and tuning these two models, the accuracy of each model is expected to be high, which contributes to the improvement of the whole turn-taking prediction.

2. A human-robot dialogue corpus

We use a human-robot dialogue corpus in which a subject talked with the android robot ERICA [34, 35] operated by another person, called an operator. The operator's voice was directly played through an audio speaker placed on ERICA. The operator also manually controlled ERICA's non-verbal behaviors such as head nodding and eye gaze. In each dialogue session, the subject was a different person whereas the operator was selected from five persons. Each dialogue session lasted about 10 minutes. This corpus consists of three dialogue tasks as below.

2.1. Job interview task

This task is a simulated job interview dialogue where ERICA was given the role of the interviewer. The operator asked the

http://dx.doi.org/10.21437/Interspeech.2019-1537

subject some questions on such as motivation of the job application and skills of the subject. Depending on the responses to the questions, the operator also made follow-up questions to elicit further information. We recorded 28 dialogue sessions with this scenario. In this task, although the interviewer (operator = ERICA) holds the dialogue initiative, the majority of utterances was done by the interviewee (subject). As the talking style is polite, it is expected that the turn-taking cue is more clear than other tasks.

2.2. Attentive listening task

In this task, the subject talked about a specific theme and ERICA listened to the talk. The operator was asked to express listening behaviors including backchanneling and questions in order to encourage the subject's talk. The theme of the talks was selected by the subject his/herself from *impressive trip* or *recent delicious food*. We recorded 19 dialogue sessions with this scenario. In this task, the subject holds both the dialogue initiative and the majority of utterances. Due to the imbalance of the amounts of utterances between the subject and the operator, the turn-taking prediction is more difficult than other tasks.

2.3. Speed dating task

The subject talked with ERICA in the setting of first-time meeting. The purpose of this dialogue task is to socialize with a person who he / she meets for the first time. Concretely, they were asked to know each other by talking about their profiles such as hobbies and food preferences. We recorded 33 dialogue sessions with this scenario. This dialogue task is mixed-initiative, and the participants exchange the turn-taking floor frequently. Furthermore, since the talking style is casual, the turn-taking prediction is also difficult.

3. Annotation of transition relevance place

We conducted annotation of TRP with the corpus explained in the previous section. At first, we summarize existing annotation in the corpus which is used for the annotation of TRP. Then, the annotation procedure of TRP is explained, followed by analyses of the annotation result.

3.1. Existing annotation

We had already annotated following labels.

Inter-pausal unit (IPU) [36]
Each utterance is segmented by pauses longer than 200 milliseconds. Note that the turn-taking prediction is done by this unit in this study.

Long utterance unit (LUU) [37]
Each utterance is also segmented by boundaries defined by the syntactic, conversational, and interactive viewpoint. In many cases, an LUU segment consists of several IPUs. It is expected that turn-taking is more related to LUU boundaries than IPU boundaries.

Dialogue act (DA)
Each LUU segment is annotated with a dialogue act (DA) label such as question, answer, and statement. In total, we used 19 kinds of labels by referring the label set defined in an earlier study [38]. Furthermore, based on the DA labels, we annotated labels of adjacency pair that represent a corresponding pair of DA labels such as question and answer. Since these annotated labels describe the functional aspect in dialogue, they are useful for annotation of TRP labels.

3.2. Annotation procedure of TRP

In this study, we asked a third-party person to annotate the labels of TRP with the following procedure. The judgment of TRP was done at an ending point of every LUU segments annotated in the current corpus. Note that the numbers of LUU segments in the corpus were 1,550, 2,351, and 4,371 in job interview task, attentive listening task, and speed-dating task, respectively. It is theoretically pointed out that ending points of LUU segments are related to TRP [39]. At an ending point of each LUU segment, the annotator was asked to judge if a current listener is able to take the turn as the next action from the viewpoint of the listener. If the annotator perceives it is possible, this place is annotated as TRP, otherwise not. The annotator was asked not to consider the subsequent dialogue content. The annotator was asked to make the judgment of TRP basically based on the transcription text in this annotation work, though they also could refer to the dialogue video with audio as supplemental materials for the judgment. We also gave the labels of DA and adjacency pairs to the annotator as supplementary information for judging of TRP. For example, when a current speaker is answering to a question and a current LUU segment is the intermediate part of the answering, the end of the current LUU segment is not judged as TRP. Theoretically, TRP is perceived by only actual participants in the dialogue. In this sense, the current annotated labels are some approximation of the true TRP.

Besides the binary labels for *TRP* and *not TRP*, there are some ambiguous situations. To deal with this problem, we made a guideline for the ambiguous situations. For example, we observed places where there was no pause gap between LUU segments though it looked TRP when we only consider its utterance text. Since there is no chance for the listener to take the turn due to no pause gap, this case was annotated as *not TRP*. Although the ambiguous places were observed with low frequency, this guideline would be important for increasing the objectivity of this annotation work.

3.3. Inter-annotator agreement

In order to investigate the objectivity of this annotation work, we asked two persons to annotate TRP labels as a preliminary trial. This trial was conducted with two dialogue sessions for each dialogue task. After this trial annotation work, we calculated Cohen's kappa score between the two annotators. As a result, the kappa scores were 0.791, 0.784, and 0.817 for job interview task, attentive listening task, and speed dating task, respectively. Since these scores meant much higher than moderate agreement (> 0.6) and around strong agreement (> 0.8), we decided to conduct the remaining annotation work with one person. In this study, the used TRP labels were given by this person.

3.4. Distribution of annotated labels

We conducted the TRP annotation with all dialogue sessions of the three dialogue tasks and then analyzed the annotated labels. Table 1 reports the distribution of the TRP labels together with those of turn-taking labels. In the turn-taking labels, *take* means turn-switch and *not take* means turn-holding. Note that the unit of the numbers in this table is the number of IPU segments which were used in the turn-taking prediction. We observe the general tendency that the number of TRP is larger than those of

Table 1: *Number of TRP and turn-taking labels*

dialogue task	TRP label		turn-taking label	
	TRP	not TRP	take	not take
job interview	950	2,695	825 (22.6%)	2,820 (77.4%)
attentive listening	1,347	3,350	965 (20.5%)	3,732 (79.5%)
speed dating	3,075	3,778	2,337 (34.1%)	4,516 (65.9%)

Table 2: *Proportion of turn-taking labels at TRP and not TRP*

dialogue task	TRP		not TRP	
	take	not take	take	not take
job interview	84.0%	16.0%	1.0%	99.0%
attentive listening	58.4%	41.6%	5.3%	94.7%
speed dating	66.7%	33.3%	7.6%	92.4%

the turn-switch (*take*). This suggests that there are many TRP in which the turn-switch did not happen as a result.

We further analyzed the proportion of turn-taking labels at TRP and not TRP, as reported in Table 2. Among the original turn-taking labels, the number of turn-holding (*not take*) was much larger than those of turn-switch (*take*) as reported in Table 1. Considering only the places annotated as *TRP*, the ratio of turn-switch became larger than those of turn-holding. On the other hand, the ratio of turn-holding became overwhelmingly larger (over 90%) in the places annotated as *not TRP*. This result supports that the TRP labels are strongly related to the turn-taking labels and also suggests that recognition of the TRP labels contributes to the turn-taking prediction. We also observed some *not TRP* samples in which the turn-switch happened. These were mainly caused by the interruption (barge-in) by the listener. Note that this is out of scope of the proposed model.

4. Proposed model

By utilizing the TRP labels in turn-taking prediction, we propose a two-step model: (1) TRP detection and (2) turn-taking prediction at TRP. Figure 2 illustrates the architecture of the proposed model. While the conventional methods directly predicted the turn-taking events, the proposed model conducts the different two decisions in turn-taking prediction so that each model is separately designed and trained. We explain these models respectively, and then the final decision process.

4.1. TRP detection

This model detects TRP at the end of IPU, based on prosodic and linguistic information of the preceding utterance. We used a hierarchical model of LSTM where each kind of feature is modeled by an individual LSTM and the outputs of those LSTMs are concatenated and fed into to a linear layer that outputs the posterior probability of the output label [29], as shown in Figure 2. The reference labels are binary corresponding to the TRP labels annotated in Section 3. The input feature set consists of prosodic features (power, F0, and their first and second orders, 6 dimensions in total), speech feature (log Mel filterbank, 40 dimensions), and linguistic feature (word2vec [40], 100 dimensions). We found the prosodic features did not contribute in our preliminary experiment in TRP detection. This is because

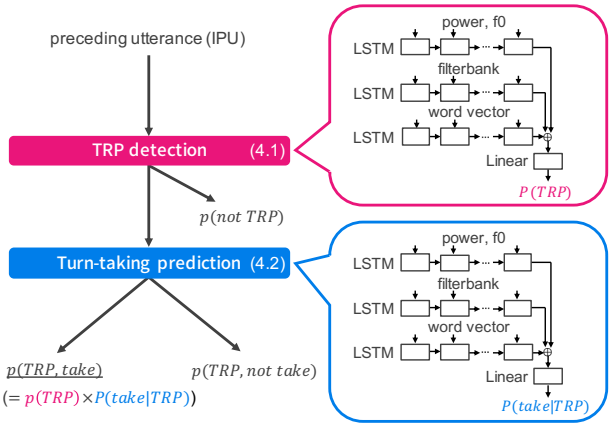

Figure 2: *Architecture of the proposed model*

the annotator made the judgment of TRP basically based on the transcription text.

4.2. Turn-taking Prediction at TRP

Given that the end of the current IPU segment is TRP, this model predicts the turn-taking label which is *take* or *not take*. We used the same model architecture and the same feature set (prosodic features, speech feature, and linguistic feature) as those of the TRP detection. The training samples are restricted to places annotated as TRP. It is expected that this prediction model is more accurate than the conventional approaches that used all samples including both *TRP* and *not TRP*.

4.3. Turn-taking prediction based on TRP detection

The proposed model predicts the turn-taking label by utilizing the above two models as follows.

1. Calculate the posterior probability of TRP $p(\text{TRP})$ by using the detection model described in Section 4.1.
2. Assuming the end of the current IPU is TRP, calculate the posterior probability of turn-switch $p(\text{take}|\text{TRP})$ by using the prediction model described in Section 4.2.
3. Using both outputs, calculate the joint probability of TRP and turn-switch as $p(\text{take}, \text{TRP}) = p(\text{TRP}) \times p(\text{take}|\text{TRP})$
4. Output the turn-switch label (*take*) if $p(\text{take}, \text{TRP}) > 0.5$, otherwise the turn-holding label (*not take*)

Note that it is assumed that it must be turn-holding (*not take*) where the place is annotated as *not TRP*, though there are some exceptional cases.

5. Experimental evaluations

The proposed model is compared with a baseline model that does not use the TRP labels and directly predict the turn-taking labels.

5.1. Setup

We conducted a 5-fold cross validation using the human-robot dialogue corpus described in Section 2. In this experiment, we predicted only the operator's (ERICA's) turn-taking behaviors. We used different strategies for training two models of TRP detection and turn-taking prediction in the proposed model. For

Table 3: *Evaluation on turn-taking prediction*

dialogue task	model	accuracy	precision	recall	F-score	F-macro
job interview	baseline	88.5	77.5	69.2	73.1	82.9
	proposed	89.5	76.3	78.1	77.2	85.2
attentive listening	baseline	80.4	54.5	26.8	36.0	62.2
	proposed	80.1	51.8	47.3	49.4	68.5
speed dating	baseline	76.3	68.2	57.0	62.1	72.4
	proposed	79.3	73.3	61.7	67.0	75.9

the TRP detection model, since the TRP labels are independent from the kind of dialogue tasks theoretically, we trained a universal model using data from all dialogue tasks. On the other hand, for the turn-taking prediction model, we trained each model for each dialogue task because the actual turn-taking behaviors depend on the kind of dialogue tasks. We used a baseline model that does not consider the TRP labels and directly predict the turn-taking labels. Its model includes LSTM layers just like the propose model. Note that the training samples are from the entire dataset, not restricted to places annotated as TRP. As evaluation metrics, we used accuracy, precision, recall, F-score, and F-macro. Accuracy is the ratio of the number of correct samples among the number of entire samples. Precision, recall, and F-score are composed for positive instances (e.g. *take*). F-macro is average of F-scores for positive and negative instances (e.g. *take* and *not take*).

The proposed and baseline models were implemented as follows. We used the PyTorch library 0.4.1 [41]. We used 3-layer LSTM with 128 nodes for the baseline model and the TRP detection model, and also utilized 1-layer LSTM with 128 nodes for the turn-taking prediction model of the proposed model. We used softmax function for the linear output layers. The loss function was cross-entropy. The minibatch size corresponded to 32 IPU segments, and model parameters were updated using RMSProp with the learning rate of 10^{-4}. The dropout was adopted to each layer with a ratio of 0.2. The input features were extracted as below. Power, pitch and log Mel filterbank were extracted in the last 2 seconds of each IPU segment with a frame shift size of 10 milliseconds. Word vectors were independently trained with continuous bag-of-words (CBOW) model by using the same training dataset.

5.2. Result

Table 3 reports the results of turn-taking prediction. Since it is important to predict the turn-switch behavior (*take*) properly in turn-taking prediction, we focus on the F-scores of the turn-switch in this evaluation. In all the dialogue tasks, the proposed model significantly improved the F-score from the baseline model, especially increased the recall rates drastically. In the baseline model, it is difficult to predict the turn-switch because the *take* label is a minor class in the entire dataset, which is imbalanced. The proposed model detects TRP and then predicts the turn-taking behavior in the restricted places so that it improved the turn-taking prediction as a whole.

We further analyzed the proposed model consisting of two steps: TRP detection and turn-taking prediction. At first, we evaluated only the TRP detection model as reported in Table 4. Since this task can be interpreted as a pre-processing to properly filter out places that are not TRP, the positive and negative instances (*TRP* and *not TRP*) are equally important. Therefore, we present the accuracy scores compared with those chance levels. The TRP detection model is able to detect TRP with accuracy of over 80% for all tasks. Next, we evaluated the turn-taking prediction model that is the second step in the proposed model. We restricted the test samples to places annotated as TRP for this evaluation of the proposed turn-taking prediction model and the baseline model. Table 5 reports F-scores together with precision and recall scores. In all the dialogue tasks, the proposed method achieved much higher F-scores, with the drastic improvement of the recall scores.

Table 4: *Evaluation on TRP detection*

dialogue task	accuracy	chance level
job interview	91.0	73.9
attentive listening	81.8	71.4
speed dating	81.7	55.1

Table 5: *Evaluation on turn-taking prediction where test samples are restricted to TRP*

dialogue task	model	precision	recall	F-score
job interview	baseline	90.0	71.4	79.7
	proposed	85.4	98.0	91.2
attentive listening	baseline	69.2	32.0	43.8
	proposed	62.6	84.6	72.0
speed dating	baseline	77.7	62.0	68.9
	proposed	71.6	88.9	79.3

6. Conclusion

We have proposed the turn-taking prediction model based on detection of TRP, which is defined as a timing when the current speaking turn can be completed and other participants are able to take the turn. We conducted manual annotation of TRP with a human-robot interaction corpus where several dialogue tasks are designed. We confirmed almost strong agreement between two persons. The proposed model consists of two neural networks: TRP detection and turn-taking prediction at TRP. At first, the model detects TRP, and if the TRP is detected, then the model predicts a turn-taking behavior. The experimental results showed that the proposed model improved turn-taking prediction from the baseline model that does not consider the TRP and directly predicts turn-taking behaviors. Furthermore, the results revealed that TRP detection was realized with accuracy of over 80% for all dialogue tasks.

7. Acknowledgment

This work was supported by JST ERATO Ishiguro Symbiotic Human-Robot Interaction Project JPMJER1401.

8. References

[1] C. Chao, J. Lee, M. Begum, and A. L. Thomaz, "Simon plays Simon says: The timing of turn-taking in an imitation game," in *RO-MAN*, 2011, pp. 235–240.

[2] G. Skantze, A. Hjalmarsson, and C. Oertel, "Turn-taking, feedback and joint attention in situated human–robot interaction," *Speech Communication*, vol. 65, pp. 50–66, 2014.

[3] S. Andrist, X. Z. Tan, M. Gleicher, and B. Mutlu, "Conversational gaze aversion for humanlike robots," in *HRI*, 2014, pp. 25–32.

[4] 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.

[5] F. Gervits and M. Scheutz, "Pardon the interruption: Managing turn-taking through overlap resolution in embodied artificial agents," in *SIGDIAL*, 2018, pp. 99–109.

[6] R. Das and H. Pon-Barry, "Turn-taking strategies for human-robot peer-learning dialogue," in *SIGDIAL*, 2018, pp. 119–129.

[7] R. Meena, G. Skantze, and J. Gustafson, "The map task dialogue system: A test-bed for modelling human-like dialogue," in *SIGDIAL*, 2013, pp. 366–368.

[8] A. Cafaro, N. Glas, and C. Pelachaud, "The effects of interrupting behavior on interpersonal attitude and engagement in dyadic interactions," in *AAMAS*, 2016, pp. 911–920.

[9] O. Niebuhr, K. Görs, and E. Graupe, "Speech reduction, intensity, and F0 shape are cues to turn-taking," in *SIGDIAL*, 2013, pp. 261–269.

[10] M. Zellers, "Pitch and lengthening as cues to turn transition in Swedish," in *INTERSPEECH*, 2013, pp. 248–252.

[11] M. Zellers, "Perception of pitch tails at potential turn boundaries in Swedish," in *INTERSPEECH*, 2014, pp. 1944–1948.

[12] A. Gravano, P. Brusco, and S. Benus, "Who do you think will speak next? Perception of turn-taking cues in Slovak and Argentine Spanish," in *INTERSPEECH*, 2016, pp. 1265–1269.

[13] P. Brusco, J. M. Pérez, and A. Gravano, "Cross-linguistic study of the production of turn-taking cues in American English and Argentine Spanish," in *INTERSPEECH*, 2017, pp. 2351–2355.

[14] Y. Ishimoto, T. Teraoka, and M. Enomoto, "End-of-utterance prediction by prosodic features and phrase-dependency structure in spontaneous japanese speech," in *INTERSPEECH*, 2017, pp. 1681–1685.

[15] A. Maier, J. Hough, and D. Schlangen, "Towards deep end-of-turn prediction for situated spoken dialogue systems," *INTERSPEECH*, 2017.

[16] I. De Kok and D. Heylen, "Multimodal end-of-turn prediction in multi-party meetings," in *ICMI*, 2009, pp. 91–98.

[17] K. Jokinen, K. Harada, M. Nishida, and S. Yamamoto, "Turn-alignment using eye-gaze and speech in conversational interaction," in *INTERSPEECH*, 2010, pp. 2018–2021.

[18] R. Ishii, K. Otsuka, S. Kumano, M. Matsuda, and J. Yamato, "Predicting next speaker and timing from gaze transition patterns in multi-party meetings," in *ICMI*, 2013, pp. 79–86.

[19] M. Roddy, G. Skantze, and N. Harte, "Multimodal continuous turn-taking prediction using multiscale rnns," in *ICMI*, 2018, pp. 78–86.

[20] R. Ishii, K. Otsuka, S. Kumano, and J. Yamato, "Analysis of respiration for prediction of who will be next speaker and when? in multi-party meetings," in *ICMI*, 2014, pp. 18–25.

[21] R. Ishii, S. Kumano, and K. Otsuka, "Analyzing mouth-opening transition pattern for predicting next speaker in multi-party meetings," in *ICMI*, 2016, pp. 209–216.

[22] M. Włodarczak and M. Heldner, "Respiratory turn-taking cues," in *INTERSPEECH*, 2016, pp. 1275–1279.

[23] M. Johansson and G. Skantze, "Opportunities and obligations to take turns in collaborative multi-party human-robot interaction," in *SIGDIAL*, 2015, pp. 305–314.

[24] J. Kane, I. Yanushevskaya, C. d. Looze, B. Vaughan, and A. N. Chasaide, "Analysing the prosodic characteristics of speech-chunks preceding silences in task-based interactions," in *INTERSPEECH*, 2014, pp. 1681–1685.

[25] N. G. Ward, O. Fuentes, and A. Vega, "Dialog prediction for a general model of turn-taking," in *INTERSPEECH*, 2010, pp. 2662–2665.

[26] 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, pp. 1661–1665.

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

[28] R. Masumura, T. Tanaka, A. Ando, R. Ishii, R. Higashinaka, and Y. Aono, "Neural dialogue context online end-of-turn detection," in *SIGDIAL*, 2018, pp. 224–228.

[29] D. Lala, K. Inoue, and T. Kawahara, "Evaluation of real-time deep learning turn-taking models for multiple dialogue scenarios," in *ICMI*, 2018, pp. 78–86.

[30] Z. Aldeneh, D. Dimitriadis, and E. M. Provost, "Improving end-of-turn detection in spoken dialogues by detecting speaker intentions as a secondary task," in *ICASSP*, 2018, pp. 6159–6163.

[31] H. Sacks, E. A. Schegloff, and G. Jefferson, "A simplest systematics for the organization of turn-taking for conversation," *Language*, vol. 50, no. 4, pp. 696–735, 1974.

[32] 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*, vol. 13, pp. 134–184, 1996.

[33] H. Tanaka, *Turn-taking in Japanese conversation: A study in grammar and interaction*. John Benjamins Publishing, 2000, vol. 56.

[34] K. Inoue, P. Milhorat, D. Lala, T. Zhao, and T. Kawahara, "Talking with ERICA, an autonomous android," in *SIGDIAL*, 2016, pp. 212–215.

[35] T. Kawahara, "Spoken dialogue system for a human-like conversational robot ERICA," in *IWSDS*, 2018.

[36] H. Koiso, Y. Horiuchi, S. Tutiya, A. Ichikawa, and Y. Den, "An analysis of turn-taking and backchannels based on prosodic and syntactic features in japanese map task dialogs," *Language and speech*, vol. 41, no. 3-4, pp. 295–321, 1998.

[37] Y. Den, H. Koiso, T. Maruyama, K. Maekawa, K. Takanashi, M. Enomoto, and N. Yoshida, "Two-level annotation of utterance-units in japanese dialogs: An empirically emerged scheme," in *LREC*, 2010, pp. 2103–2110.

[38] H. Bunt, J. Alexandersson, J. Carletta, J.-W. Choe, A. C. Fang, K. Hasida, K. Lee, V. Petukhova, A. Popescu-Belis, L. Romary, C. Soria, and D. Traum, "Towards an iso standard for dialogue act annotation," in *LREC*, 2010, pp. 2548–2555.

[39] H. Koiso and Y. Den, "Towards a precise model of turn-taking for conversation: A quantitative analysis of overlapped utterances," in *DiSS-LPSS Joint Workshop*, 2010, pp. 55–58.

[40] T. Mikolov, K. Chen, G. Corrado, and J. Dean, "Efficient estimation of word representations in vector space," in *ICLR*, 2013.

[41] A. Paszke, S. Gross, S. Chintala, G. Chanan, E. Yang, Z. DeVito, Z. Lin, A. Desmaison, L. Antiga, and A. Lerer, "Automatic differentiation in PyTorch," in *NIPS Wrokshop*, 2017.