Interspeech 2018
2-6 September 2018, Hyderabad

Prediction of Turn-taking Using Multitask Learning with Prediction of Backchannels and Fillers

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

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

Abstract

We address prediction of turn-taking considering related behaviors such as backchannels and fillers. Backchannels are used by the listeners to acknowledge that the current speaker can hold the turn. On the other hand, fillers are used by the prospective speakers to indicate a will to take a turn. We propose a turn-taking model based on multitask learning in conjunction with prediction of backchannels and fillers. The multitask learning of LSTM neural networks shared by these tasks allows for efficient and generalized learning, and thus improves prediction accuracy. Evaluations with two kinds of dialogue corpora of human-robot interaction demonstrate that the proposed multitask learning scheme outperforms the conventional single-task learning.

Index Terms: turn-taking, backchannel, filler, neural networks, multitask learning

1. Introduction

In the past years, a variety of spoken dialogue systems have been deployed in smartphones, smart speakers, and humanoid robots. In a majority of the applications, the systems are designed to conduct specific tasks or information retrieval such as queries on weathers, public transportation, and personal schedules. In these scenarios, users are expected to utter a query of one sentence, which will be responded by the system. In this kind of dialogue, turn-taking protocol is explicit to the user. Actually, the push-to-talk protocol is used in smartphones, and magic words are used in smart speakers.

On the other hand, this turn-taking protocol is not natural in human-human conversations, in which many sentences are uttered in a turn, and backchannels are occasionally generated during the dialogue partner's turn. Advanced studies on spoken dialogue systems should include human-like dialogue in terms of tasks and styles. When we adopt humanoid robots or human-like agents, users naturally expect the system to talk like a human. The tasks of the system are extended to cover long conversations including job interviews and attentive listening to senior people. In this kind of dialogue, the systems are required to conduct human-like turn-taking behaviors.

Turn-taking is actually a very important and difficult problem in spoken dialogue systems, and that is a reason why current smartphones and smart speakers adopt the push-to-talk interface or magic words. Without these constraints, inappropriate turn-taking easily leads to crash of dialogue, for example, the system starts speaking while the user keeps or takes a turn. Conventional systems often assume a long pause (e.g. 1 second) implies the end of a turn. But this is not only natural but also may trigger the user's next utterance, which will be crashed by the system's utterance. Therefore, the problem of appropriate turn-taking is reduced to detection of the end of the user turns as early as possible based on the content and prosodic cues of his/her utterance. In this paper, we focus on the prosodic cues because automatic speech recognition of spontaneous conversations by distant microphones deployed in a humanoid robot is still difficult and produces many errors.

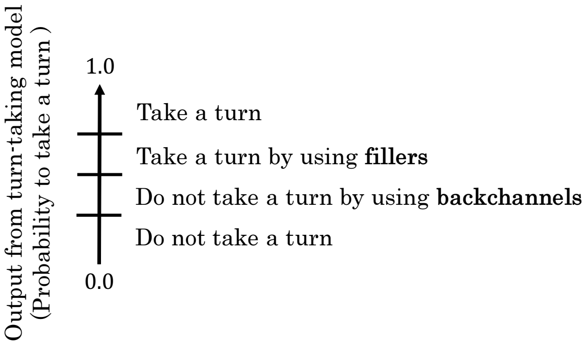
Output from turn-taking model
(Probability to take a turn)
1.0
Take a turn
Take a turn by using **fillers**
Do not take a turn by using **backchannels**
Do not take a turn
0.0

Figure 1: *Possible turn-taking behaviors considering relationship with backchannels and fillers*

Detection of the end of turns (with prosodic cues) may not be so easy even for humans. In fact, the end of turns is defined by turn-taking by the dialogue partner; if the dialogue partner does not take a turn, the current speaker might continue the turn. When the turn-switching/keeping is ambiguous, we can use backchannels or fillers. When we do not want to take a turn, we can generate backchannels to encourage the dialogue partner to keep talking. On the other hand, fillers are used to suggest to take a turn. Thus, we can make a conceptual plot of turn-taking probability and possible actions in Figure 1.

It is well-known that backchannels and fillers are related with the turn-taking behavior [1, 2], and prediction of these events, particularly backchannels, has been intensively studied using many kinds of features and machine learning techniques [3, 4]. Recently, neural network models such as LSTM are introduced to the problem of turn-taking [5, 6]. However, these problems are separately formulated and the prediction models are independently trained.

In this paper, we address joint training of turn-taking and prediction of backchannels and fillers by considering their relationship. Since these problems share the feature extraction process, we can design a neural network with shared layers and multitask learning. The proposed integrated model consists of a shared LSTM layer and individual feed-forward layers to determine the next action on not only turn-taking but also backchannels and fillers simultaneously. This joint learning will lead to a synergy in the prediction performance. Moreover, the framework will realize human-like turn-taking behaviors, for example, to generate a filler before taking a turn.

10.21437/Interspeech.2018-1442

2. Related Works

2.1. Turn-taking Prediction

There are many studies to analyze turn-taking behaviors and predict turn-taking at the end of the utterances by using features of the preceding utterances and to decide whether the current speaker's turn has ended. A large majority of them showed that prosodic features such as pitch and intensity provide useful cues [7, 8, 9, 10, 11], and many studies conducted prediction using machine learning such as SVM and neural networks [2, 12, 1].

There are some works that investigated other features in speech such as N-gram model [13], dependency structures [14], and the previous turn-taking behaviors [15]. There are also other works that investigated non-verbal features such as respiratory features [16], head pose features [17], and eye-gaze features [18].

In this paper, we focus on prosodic features which are obtained easily and robustly. In the past works listed above, a variety of parameterization methods were investigated, but very recently, a simple recurrent neural network model or LSTM has been introduced to input frame-wise pitch and intensity values and process in sequence to extract effective features to minimize the objective function, in this case, errors of prediction of turn-taking events [5, 6, 19].

While Masumura et al. [6] and Liu et al. [19] also incorporated word embedding information, we focus on prosodic features because the primary goal of this work is to investigate the joint learning of turn-taking and backchannel/filler prediction.

2.2. Backchannel Prediction

A backchannel is a short response such as "*um*" and "*right*" which the listener utters without taking the speaker's turn. Backchannels have a function to encourage the current speaker to hold the turn and continue to speak. There are a number of studies to predict backchannels by using prosodic features of the preceding utterance such as pitch and power [20, 21, 22, 23, 3], as well as linguistic features [24, 25]. Note that the prosodic features are generally same as those used in turn-taking prediction addressed in the previous sub-section.

2.3. Filler Prediction

A filler is a short phrase which fills a pause in conversations, such as "*uh*" and "*so*". Fillers have a function to indicate that the interlocutor is thinking about the next utterance and to relieve an embarrassed silence. Thus, fillers suggest holding the current turn, or taking a turn.

Several studies investigated the prosodic features of fillers [26, 27], but there are only a limited studies on prediction of fillers [28, 29].

3. Dialogue Corpora

We have collected corpora of several dialogue tasks between subjects and an android (humanoid robot) ERICA [30]. ERICA looks and behaves like a female including facial expressions and nodding, but was remotely operated by a female actress for the corpus collection. As a result, the collected dialogue was very close to human-human dialogue. In this study, the following two kinds of dialogue corpora are used.

Job interview
ERICA conducts a job interview on a subject, who plays a role of an applicant to some company which he/she was interested in. An interview was done by asking questions regarding the motivations and the skill of the applicant. There are 15 sessions and each session is about 10 minutes long. There are four operators of ERICA, one assigned to each session. Subjects are college students and differ from one session to another. In this task, the interviewer (robot operator) has an initiative of the dialogue, but a majority of the utterances are done by the interviewee (subject). Backchannels are generated mostly from the interviewer in this task.

Table 1: *Detail of prediction points after IPU. In this table,* ***BC*** *means an occurrence of backchannels and* ***NOT*** *means not occurring of backchannels, and* ***Filler*** *means an occurrence of fillers and* ***NOT*** *means not occurring of fillers.*

Corpus	Preceding speaker	Prediction points	
		SWITCH	**KEEP**
Job interview	Interviewer	175	642
	Interviewee	**276**	**620**
Attentive listening	**Speaker**	**306**	**1,131**
	Listener	162	364

Corpus	Interlocutor	Prediction points	
		BC	**NOT**
Job interview	**Interviewer**	**223**	**884**
	Interviewee	52	894
Attentive listening	Speaker	74	626
	Listener	**741**	**1,175**

Corpus	Interlocutor	Prediction points	
		Filler	**NOT**
Job interview	**Interviewer**	**473**	**1,580**
	Interviewee	287	1,766
Attentive listening	Speaker	403	2,213
	Listener	**286**	**2,330**

Attentive listening
ERICA conducts an attentive listening to an elderly subject. The subject talks about some topic he/she chose beforehand, such as "memorable travel" and "important life event". An attentive listening was done by using backchannels and some questions. There are 15 sessions and each session is about 10 minutes long. There are four operators of ERICA, one assigned to each session. Subjects differ from one session to another. In this task, the speaker (subject) has an initiative of the dialogue, and makes a large majority of the utterances in the session. There are a few turn-taking events by the listener (robot operator), thus their prediction is very difficult. On the other hand, she uses many backchannels.

Since the turn-taking behaviors differ depending on the task and initiative of the dialogue, we train a dedicated model to each corpus though the model architecture is the same. Moreover, the turn-taking behaviors differ between the robot and subjects in these tasks. Thus, we train separated models for the subjects and for the robot, but we focus on the behavior of the robot (dialogue system) for evaluation.

Prediction of turn-taking, backchannels and fillers are conducted after each utterance (IPU) by the subject. The statistics of the prediction points are listed in Table 1.

4. Integrated Prediction Model by Multitask Learning

4.1. Baseline Model

In this study, the turn-taking model is built referring to the LSTM-based model proposed by Skantze [5]. This model is prepared for each interlocutor role. This model receives the features described below at each time frame, and it outputs the probability that the target interlocutor will be speaking within 1 sec. The frame shift size is 50 msec. Therefore, an output is a 20-dimensional vector (=1000 msec / 50 msec) and each dimensional value corresponds to the utterance probability at each time frame in the future. Then, the label used for training the model has also 20 dimensions and each dimension is binary, which represents whether the interlocutor utters at the time frame. Input features are voice activity, pitch, intensity and spectral stability, and these are extracted for each interlocutor.

Voice activity: A binary feature representing the current voice activity (speech/no speech) of each interlocutor. The voice activity was extracted from the annotation of the corpus.

Pitch: The pitch and Δ, ΔΔ value of it were automatically extracted by using Praat[1] and then z-normalized for the individual interlocutors.

Intensity: The intensity and Δ, ΔΔ value of it were automatically extracted by using Praat, and then z-normalized for the individual interlocutors.

Spectral stability: The spectral stability was calculated by

$$S_t = \frac{\sum_{f=1}^{N} \mathrm{abs}(s_{t,f} - s_{t-1,f})}{\sum_{f=1}^{N} s_{t,f}} \qquad (1)$$

where N is the number of frequency bins and $s_{t,f}$ is the power in frequency bin f at time t. This value was z-normalized for the individual interlocutors. The power spectrum was extracted by using *librosa*, which is one of Python libraries.

These 16-dimensional features are input to LSTM sequentially frame by frame. The baseline model has one LSTM (18 units) and three fully connected layers (20 units in these three layers) are added prior to the output.

In this paper, turn-taking is predicted using this model at the end of each IPU. IPU is detected when a 200 msec pause goes by after the end of an utterance. This pause length was decided to surely judge whether each utterance is IPU or not. If this length is too short like 50 msec, the prediction point may be where prediction is not necessary, for example, obviously during the turn. If this length is too long like 500 msec, it is too late to predict turn-taking. At each prediction point, the sum of 20-dimensional output is calculated for each interlocutor, then the interlocutor with a larger value is judged as the next speaker. **SWITCH** is defined as the result of prediction when the preceding speaker and the next speaker are not the same people, and **KEEP** is when the preceding speaker and the next speaker are same. We exclude points from evaluation when either interlocutor starts speaking before 200 msec pause from the end of each utterance or when both interlocutors will speak or will not speak within future 1 sec.

4.2. Integrated Model

The model is extended to an integrated model which predicts not only turn-taking but also backchannels and fillers. Figure 2 depicts the proposed model in this study. In this model, the layers shared by the prediction tasks extract the common features, and separated layers for each prediction task makes a decision. This makes it possible to consider the relationship between the different tasks and to better predict turn-taking at points when it is difficult to predict. The same 16-dimensional features as the baseline model are used. This model is optimized by the following objective function.

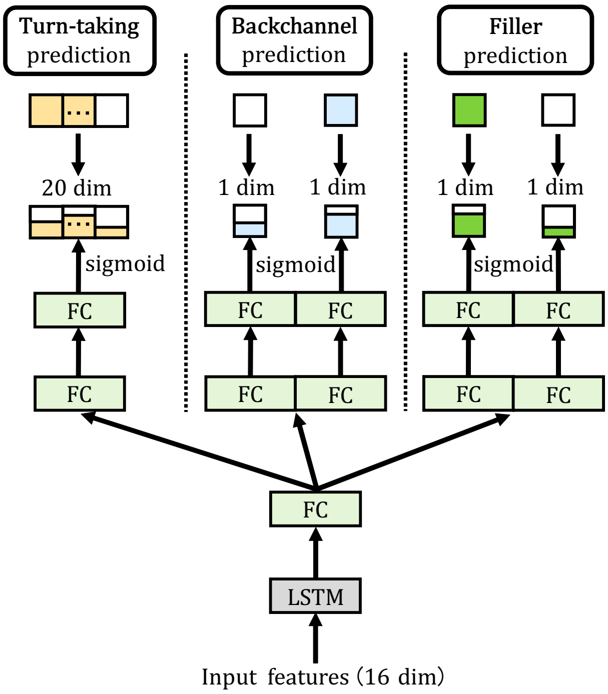

Figure 2: *Integrated model utilizing multitask learning (FC: Fully Connected layer). Backchannel and filler are predicted for each interlocutor.*

$$\mathcal{L} = \alpha \times \mathcal{L}_{turn} + \beta \times (\mathcal{L}_{bc} + \mathcal{L}_{filler}) \qquad (2)$$

where $\mathcal{L}_{turn}$ is a loss function for the turn-taking prediction, $\mathcal{L}_{bc}$ and $\mathcal{L}_{filler}$ are loss functions for the backchannel and filler prediction. α and β represent a weight for each loss function. In addition, the loss function for each prediction task was calculated by the following function.

$$\mathcal{L}_{task} = \begin{cases} r_{task,N} \times \mathrm{MSE}_{task} & \text{(if positive instance)} \\ r_{task,P} \times \mathrm{MSE}_{task} & \text{(otherwise)} \end{cases} \qquad (3)$$

where $r_{task,N}$ is the rate of negative instances in $task$ and $r_{task,P}$ is the rate of positive instances in $task$. MSE_{task} is a mean-squared error loss function for $task$.

In order to predict backchannels and fillers, this model outputs probabilities whether the event occurs within 1 sec. in the future. Therefore, the label is a binary value. For prediction of backchannel/filler prediction, we use the same points as those for turn-taking. Since backchannels are not used by the current speaker at the end of the utterance, such points are excluded from backchannel prediction.

[1] Boersma, Paul & Weenink, David: Praat, http://www.praat.org/ (2016).

Table 2: *Result of turn-taking prediction* **after subject's (interviewee's) utterance** *in* **the job interview corpus**

model	SWITCH			KEEP		
	Precision	Recall	F score	Precision	Recall	F score
baseline	79.3	69.2	73.9	87.0	91.9	89.4
multitask	84.2	71.4	77.3	88.1	94.0	91.0

Table 3: *Result of turn-taking prediction* **after subject's (speaker's) utterance** *in* **the attentive listening corpus**

model	SWITCH			KEEP		
	Precision	Recall	F score	Precision	Recall	F score
baseline	54.6	48.7	51.5	86.5	89.0	87.8
multitask	60.4	47.4	53.1	86.5	91.6	89.0

5. Experimental Evaluations

5.1. Setup

Evaluations for each dialogue corpus were done with a 5-fold cross validation. Prediction models are trained for each interlocutor role in each dialog corpus (e.g. interviewer (robot) and interviewee (subject) in the job interview corpus).

A sigmoid activation function was used for the output layers and a ReLU function was used for the hidden layers. The weights were optimized for each mini-batch including 30 samples (a sample is 20 sec. time series data). They were updated using RMSProp and the learning rate was set to 1.6×10^{-3}. The learning rate was halved when the loss for the test data was not reduced during 10 epochs, and it was halved up to 4 times at most. The dropout ratio between each layer was set to 0.2. We used 18 hidden nodes in the LSTM layer and 20 hidden nodes in the upper fully connected layers.

For the weights of multitask learning in Equation (2), we used $\alpha = 0.6$ and $\beta = 0.2$ (for each of 2 sub-tasks) when turn-taking prediction is the main task. When turn-taking prediction is not the main task, we used 0.8 as the weight of the main task and 0.1 as the weights of two predictions of turn-taking (subject's turn prediction and robot's turn prediction). All networks were implemented with Keras 2.1.2 [2] using TensorFlow 1.4.1 as backend. We adopted precision, recall and F-score as evaluation measures for each prediction task.

5.2. Results

Since we were interested in the robot's (system's) turn-taking behavior after the subject utterance, Table 2 and 3 show the results of turn-taking prediction when the preceding speaker is the subject. It is observed that the **SWITCH** precision of both corpora were significantly improved by the multitask learning. Considering the information of not occurring backchannels and occurrence of fillers by the next speaker, the proposed model could predict turn-taking more accurately even when it is difficult to predict. It is shown that the **KEEP** recall of both corpora were also improved by the multitask learning. This is because the proposed model can consider the occurrence of backchannels of the current listener and fillers of the current speaker.

Prediction results of backchannels and fillers are reported in Table 4, 5 and 6 respectively. In the job interview corpus, the multitask learning improved both of the backchannel prediction and the filler prediction. The result indicates that the robot's (interviewer's) turn-taking behavior (especially at the points where the turn-taking is ambiguous) is related to the prediction of backchannels and fillers. In the attentive listening corpus, prediction of fillers is not conducted because they are not frequent. The multitask learning improved the recall of the backchannel prediction in this corpus, too. The result suggests that the robot's (listener's) turn-taking behavior is related to the backchannel prediction, because the listener encourages the speakers' utterances by using many backchannels.

Table 4: *Result of* **robot's (interviewer's) backchannel** *prediction in* **the job interview corpus**

model	BC		
	Precision	Recall	F score
baseline	25.8	79.4	38.9
multitask	27.8	87.0	42.1

Table 5: *Result of* **robot's (listener's) backchannel** *prediction in* **the attentive listening corpus**

model	BC		
	Precision	Recall	F score
baseline	45.5	89.2	60.3
multitask	44.2	91.4	59.6

Table 6: *Result of* **robot's (interviewer's) filler** *prediction in* **the job interview corpus**

model	Filler		
	Precision	Recall	F score
baseline	31.4	84.4	45.8
multitask	31.1	88.8	46.0

6. Conclusions

In this paper, we have proposed a turn-taking prediction model considering related behaviors of backchannels and fillers. This model is based on the shared network structure and multitask learning. It was trained for each role of different dialogue corpora considering the different characteristics of dialogue behaviors. The experimental evaluations demonstrated that the proposed model predicted turn-taking better than the single-task model in the two kinds of corpora.

7. Acknowledgments

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

8. References

[1] 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.

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

[3] M. Mueller, D. Leuschner, L. Briem, M. Schmidt, K. Kilgour,

[2] Chollet, F. et al.: Keras, https://github.com/keras-team/keras (2015).

S. Stueker, and A. Waibel, "Using neural networks for data-driven backchannel prediction: A survey on input features and training techniques," in *HCI International*, 2015, pp. 329–340.

[4] T. Kawahara, T. Yamaguchi, K. Inoue, K. Takanashi, and N. G. Ward, "Prediction and generation of backchannel form for attentive listening systems," in *INTERSPEECH*, 2016, pp. 2890–2894.

[5] G. Skantze, "Towards a general, continuous model of turn-taking in spoken dialogue using LSTM recurrent neural networks," in *SIGdial*, 2017, pp. 220–230.

[6] R. Masumura, T. Asami, R. Masataki, and R. Higashinaka, "Online end-of-turn detection from speech based on stacked time-asynchronous sequential networks," in *INTERSPEECH*, 2017, pp. 1661–1665.

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

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

[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] 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.

[11] 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.

[12] J. Kane, I. Yanushevskaya, C. de Looze, B. Vaughan, and A. N. Chasaide, "Analysing the prosodic characteristics of speech-chunks preceding silences in task-based interactions," in *INTERSPEECH*, 2014, pp. 333–337.

[13] L. Ferrer, E. Shriberg, and A. Stolcke, "Is the speaker done yet? Faster and more accurate end-of-utterance detection using prosody," in *ICSLP*, 2002, pp. 2061–2064.

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

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

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

[18] 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.

[19] 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.

[20] N. G. Ward, "Using prosodic clues to decide when to produce back-channel utterances," in *ICSLP*, vol. 3, 1996, pp. 1728–1731.

[21] N. G. Ward and W. Tsukahara, "Prosodic features which cue back-channel responses in English and Japanese," *Journal of pragmatics*, vol. 32, no. 8, pp. 1177–1207, 2000.

[22] S. Fujie, K. Fukushima, and T. Kobayashi, "Back-channel feedback generation using linguistic and nonlinguistic information and its application to spoken dialogue system," in *INTERSPEECH*, 2005, pp. 889–892.

[23] K. P. Truong, R. Poppe, and D. Heylen, "A rule-based backchannel prediction model using pitch and pause information," in *INTERSPEECH*, 2010.

[24] 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.

[25] N. Kitaoka, M. Takeuchi, R. Nishimura, and S. Nakagawa, "Response timing detection using prosodic and linguistic information for human-friendly spoken dialog systems," *Transactions of the Japanese Society for Artificial Intelligence*, vol. 20, no. 3, pp. 220–228, 2005.

[26] M. Watanabe, *Features and roles of filled pauses in speech communication: A corpus-based study of spontaneous speech*. Hituzi Syobo, 2009.

[27] S. Nakamura, R. Nakanishi, K. Takanashi, and T. Kawahara, "Analysis of the relationship between prosodic features of fillers and its forms or occurrence positions," in *INTERSPEECH*, 2017, pp. 1726–1730.

[28] S. Andersson, K. Georgila, D. Traum, M. Aylett, and R. A. Clark, "Prediction and realisation of conversational characteristics by utilising spontaneous speech for unit selection," in *Speech Prosody*, 2010.

[29] R. Nakanishi, K. Inoue, K. Takanashi, and T. Kawahara, "Generating fillers based on dialog act pairs for smooth turn-taking by humanoid robot," in *IWSDS*, 2018.

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