Generating Fillers based on Dialog Act Pairs for Smooth Turn-Taking by Humanoid Robot

Ryosuke Nakanishi, Koji Inoue, Shizuka Nakamura, Katsuya Takanashi, and Tatsuya Kawahara

Abstract In spoken dialog systems for humanoid robots, smooth turn-taking function is one of the most important factors to realize natural interaction with users. Speech collisions often occur when a user and the dialog system speak simultaneously. This study presents a method to generate fillers at the beginning of the system utterances to indicate an intention of turn-taking or turn-holding just like human conversations. To this end, we analyzed the relationship between a dialog context and fillers observed in a human-robot interaction corpus, where a user talks with a humanoid robot remotely operated by a human. At first, we annotated dialog act tags in the dialog corpus and analyzed the typical type of a sequential pair of dialog acts, called a DA pair. It is found that the typical filler forms and their occurrence patterns are different according to the DA pairs. Then, we build a machine learning model to predict occurrence of fillers and its appropriate form from linguistic and prosodic features extracted from the preceding and the following utterances. The experimental results show that the effective feature set also depends on the type of DA pair.

Ryosuke Nakanishi
School of Informatics, Kyoto University, Japan, e-mail: nakanisi@sap.ist.i.kyoto-u.ac.jp

Koji Inoue
School of Informatics, Kyoto University, Japan, e-mail: inoue@sap.ist.i.kyoto-u.ac.jp

Shizuka Nakamura
School of Informatics, Kyoto University, Japan, e-mail: shizuka@sap.ist.i.kyoto-u.ac.jp

Katsuya Takanashi
School of Informatics, Kyoto University, Japan, e-mail: takanasi@sap.ist.i.kyoto-u.ac.jp

Tatsuya Kawahara
School of Informatics, Kyoto University, Japan, e-mail: kawahara@i.kyoto-u.ac.jp

1 Introduction

A number of spoken dialog systems have been developed and used with smartphones and appliances. The majority of these systems assume that the user utters a query made of single sentence, to which the system responds. In these systems, turn-taking is explicit; the user explicitly signals the start of the utterance with the "push-to-talk" interface or a predefined magic word, and the system indicates when it can accept input with an LED or GUI. However, this is much different from the turn-taking manner in human-human dialog, and is difficult to be applied for humanoid robots designed to be engaged in natural interaction with human. The goal of this study is to realize natural conversational behavior of an autonomous android [7] including backchannels [11] and turn-taking [20].

In fact, without the explicit interfaces mentioned above, speech collisions often occur when a user and the dialog system speak simultaneously, and they usually result in speech recognition errors and dialog breakdown. For natural and smooth turn-taking, this study investigates generation of fillers at the beginning of the system utterances to indicate an intention of turn-taking or turn-holding. For example, by placing a filler, the current speaker can hold the turn while thinking the next utterance, or the other participant can take a turn smoothly before speaking the main utterance. Thus, fillers have an important role and effect in human-human conversations.

In spoken language processing, however, fillers have been regarded as redundant segments, which must be removed in the transcript and are not usually generated by the system. There are several studies to predict fillers [1], but their major aim is to detect and remove them in the speech input. While there are a number of works on prediction and generation of backchannels [11], there are only a limited trials on filler generation [20, 19, 2, 21]. Fillers have a variety of forms especially in Japanese, and they are used in different contexts. Watanabe [22] investigated the occurrence ratio of fillers based on the complexity of the following utterance and the syntactic boundary. In this work, we present a method to predict occurrence of fillers and its appropriate form based on dialog act pairs using the linguistic and prosodic features in adjacent utterances. In the remaining of the paper, we describe the corpus and the annotation in Section 2, analysis on dialog act pairs and fillers in Section 3. Prediction of fillers and its evaluation is conducted in Section 4.

2 Corpus and annotation

2.1 Corpus

We use a human-robot interaction corpus, where a subject talks with an android ERICA [15, 14] remotely operated by a human operator, who talks with a subject and controls non-verbal behaviors. The recording was done in September 2016.

There were 39 sessions and each session lasted around 10 minutes. Engaged in these sessions were 6 operators (female from 20's to 30's) and 39 subjects (male from 20's to 60's and female from 10's to 70's). The android was given a role of a laboratory secretary and the subjects were asked to talk with her as a visitor.

2.2 Annotation

We define fillers as filled pauses which is one of spontaneous speech disfluency, such as "*uh, um, oh*" in English and "*ano:, etto, ma*" in Japanese [13]. For dialog act (DA) annotation, we adopt the following simple four classes based on the classification of general-purpose functions proposed by Bunt et al. [3].

- *Question* (Q): Utterances which function as information-seeking
- *Statement* (S): Utterances which have a role of *Inform* / *Offer* / *Promise* / *Request* / *Instruct*
- *Response* (R): Utterances which respond to a specific DA such as *Answer*, *Accept Offer* and *Decline Offer*
- *Other* (O): Utterances which do not belong to either Q, R or S such as *Greeting* and *Apology*

Utterances corresponding to *Feedback* in the dimension-specific function among those of O were classified into R because a number of those are lexical responses such as "*I see*". Backchannels are excluded from this annotation. To define the unit of DA, we adopt long utterance units (LUU) [6] which are defined as syntactic and pragmatic disjuncture based on clause units.

In order to validate the reliability of the DA annotation (Q, R, S and O), we calculate Cohen's kappa value [4] and evaluate agreement between annotators. The obtained kappa value was high (κ=0.799) from the result of two annotators engaged on three sessions (495 DAs).

2.3 Annotation Results

The total number of DA units was 6441 in all 39 sessions. The occurrence counts of Q, R, S and O by the operators are 758, 1064, 779 and 706, and the occurrence counts of Q, R, S and O by the subjects are 267, 1687, 477 and 703, respectively. In the following, we do not distinguish DAs by the role.

Then, the concept of adjacency pairs of DA units [18] is introduced. We extracted 5080 DA pairs by eliminating DA pairs with an overlap between the preceding DA and the following DA. The total number of fillers observed in the corpus is 4292. In this study, we focus on 1460 fillers (Operator's: 875, Subject's: 585) which occur at the beginning of the following utterance. It is presumed that these fillers are related with the turn management. Table 1 shows the ratio of DA pairs which have a filler

Table 1 Ratio of filler occurrence within DA pairs

DA pairs in turn-switch	33.2% (836/2516)
DA pairs in turn-keep	24.3% (624/2564)
Total DA pairs	28.7% (1460/5080)

between the preceding DA and the following DA. It is observed that fillers are more likely to be used in turn-switching or turn-taking than in turn-keeping cases.

3 Analysis based on dialog act (DA) pairs

3.1 Typical DA pairs and possible speech collisions

Table 2 shows the bigram statistics of DA pairs with classification of turn-holding/taking. Each column represents the preceding DA and each row represents the following DA. Turn-holding means both the preceding DA and the following DA are spoken by the same speaker, and turn-taking means the speaker of the following DA differs from the speaker of the preceding DA. Bigrams are normalized so that the sum over each preceding DA (=column) becomes 1. The numbers in parentheses indicate the occurrence counts of DA pairs. Bigram patterns with frequencies larger than 152 (3% of all DA pairs) are written in bold, and these are focused in the analysis and prediction experiment in this study. The DA pairs related with Other (O) are excluded because most of the utterances tagged with O are the typical expressions like "*sorry*" and the turn-taking behavior in these cases should be different. There are many patterns of typical DA pairs in the corpus, due to a mixed-initiative dialog characteristic between a secretary and a visitor.

First, it is confirmed in the first row that, after questions (Q) by one participant, responses (R) by the other participant are dominant. In this case, there should be a consensus of turn-switching by the both participants. However, as we see in the second column, after responses (R), there is much ambiguity in the following DA. There is ambiguity in turn-management, too. The same participant can continue his utterances (R or S), or the other participant can take back a turn. In the latter case, he/she can either respond to the previous utterance (R) or ask a new question (Q). After statements (S), there is not a large variation in the following DA, but big ambiguity in turn-management. The same speaker can continue a statement (S) or the other participants make a response (R) to the statement. It is expected that speech collisions are likely to occur when it is ambiguous which participant should take a turn. The above-mentioned cases are typical. In summary, speech collisions will often occur in the transition from R or S.

Table 2 Bigrams of DA

			Preceding DA Q	R	S	O
Following DA	Turn-holding	Q	0.09 (75)	**0.08 (187)**	0.09 (97)	0.11 (108)
		R	0.01 (6)	**0.33 (731)**	0.00 (5)	0.03 (26)
		S	0.04 (30)	**0.12 (266)**	**0.34 (365)**	0.20 (197)
		O	0.02 (19)	0.07 (146)	0.06 (69)	0.24 (237)
	Turn-taking	Q	0.02 (15)	**0.12 (257)**	0.08 (82)	0.06 (64)
		R	**0.79 (627)**	**0.15 (336)**	**0.26 (274)**	0.02 (22)
		S	0.01 (11)	**0.07 (164)**	0.06 (59)	0.07 (71)
		O	0.02 (14)	0.06 (134)	0.11 (117)	0.27 (269)

3.2 Typical DA pairs and filler patterns

Next, we investigate typical filler patterns for each of the DA pairs which are identified as important in the previous sub-section. We classify fillers into six classes defined by the function and the expression similarity as shown in Table 3.

The typical forms and their occurrence ratio for DA pairs focused in this study are shown in Table 4. We focus on the filler occurrence or not between the preceding DA and the following DA. The upper part and the lower part show the DA pairs of *Keep* (turn-holding) and *Switch* (turn-taking), respectively. The table also gives the most typical or dominant form of fillers used in the DA pair, and then the frequency ratio of that form and other forms as well as no-filler occurrence.

As shown in the upper parts, when the speaker tries to hold the turn, he does not use the notice form, but demonstrative or proper forms. They are used to hold the turn to take time before speaking the next utterance. And the ratio of no-filler occurrence is large in turn-keeping. On the other hand, the lower part suggests that when the speaker is changed, the notice form is most frequently used. The form indicates a response to the preceding utterance and causes natural turn-switching. In QR *Switch* where the turn-switching is most apparent, the ratio of filler usage is smaller and the notice form is not so dominant. It is found out that the tendency of filler usage and its typical form is different depending on the DA pair and in particular turn-switching. Based on this observation, we design prediction and generation of fillers.

Table 3 Filler class and its definition

Class and ratio of occurrence	Definition
proper (p) 7%	forms only used as fillers (“*um*” in English / “*etto*” in Japanese)
demonstrative (d) 6%	same forms as demonstrative adjectives (“*so*” in English / “*ano*” in Japanese)
adverbial (a) 2%	same forms as adverbs (“*well*” in English / “*ma-*” in Japanese)
notice (n) 12%	used to indicate a reaction (“*oh*” in English / “*a*” in Japanese)
no filler (nf) 71%	not to generate fillers

Table 4 Typical form of each DA pair

DA pair	Typical Form	Ratio of occurrence (tf[1]/ o[2]/ nf[3])
RQ Keep	proper	14% / 13% / 73%
RR Keep	demonstrative	12% / 12% / 76%
RS Keep	demonstrative	11% / 21% / 68%
SS Keep	demonstrative	11% / 14% / 75%
QR Switch	**notice**	16% / 17% / 67%
RQ Switch	**notice**	24% / 17% / 59%
RR Switch	**notice**	39% / 2% / 59%
RS Switch	proper	10% / 20% / 70%
SR Switch	**notice**	30% / 7% / 63%

[1, 2, 3] typical form, other fillers, and no filler, respectively

4 Prediction of fillers

4.1 Category for prediction

We assume that the DA of the previous utterance (of either participant) is given and the DA of the next utterance (either holding or taking the turn) is determined. Then, we want to predict the occurrence of a filler and its form. Since there is a different typical form depending on the DA pair as shown in the previous section, the filler form to be predicted is limited to the typical form and other forms collectively. Thus, the target of prediction is reduced to typical form, other forms (o) and no-filler (nf). Moreover, if the ratio of the other forms is small (smaller than 10%) or the number of filler samples is small, they are merged into the typical filler making a single filler category (f).

The baseline single model to predict only occurrence of fillers is also trained by using all data without consideration of DA pairs. This single model outputs the typical filler depending on the DA pair if it predicts filler occurrence.

4.2 Classifier and features

We use the Random Forest classifier from Scikit-learn [17] and evaluate its performance in 5-fold cross-validation. The number of decision trees which are built by bootstrap is set to ten. Since the number of samples is very different according to the class, we conduct normalization by sub-sampling in training, but the evaluation via cross-validation is conducted for the entire set. The evaluation measures are precision, recall and F-measure (their harmonic mean).

We incorporate linguistic (L) and prosodic (P) features extracted from the preceding utterance (pLUU: preceding LUU) and linguistic (L) features extracted from the following utterance (fLUU: following LUU). They are listed in Table 5.

We adopt features of last words and the boundary type as Japanese has a characteristic in the end-of-sentence expressions [5]. As the length of the preceding utterance is also an important feature, we include the number of words and chunks. Moreover, prosodic features are related with the distinction between turn-taking and turn-holding [12] [16]. Therefore, we extract them from the end of the preceding utterance. F0 and power are extracted by STRAIGHT [9, 10] (XSX [8]). We calculate the regression coefficient, the mean value, the maximum value, and the minimum value of F0 and power. The speech rate is approximately computed by dividing the number of characters by the duration, and the pause is defined as the time (ms) from the end of the pLUU.

The feature extracted from the following utterances (fLUU) would be useful for filler prediction, but we need to conduct prediction before the following utterance. In fact, people generate fillers while thinking of the next utterances. It is unrealistic to get exact information of the following utterance. In this work, however, we assume that the beginning word and the approximate length (two classes of short or long) is determined, and these features are used for prediction. The assumption of availability of these features will stand at least for the system side, and the goal of this study is generation of fillers by the robot after deciding what to speak.

4.3 Prediction performance

Table 6 shows the prediction result (F-measure) for each DA pair. We tested all possible combinations of three feature sets (pLUU(L), pLUU(P) and fLUU(L)) and the results with the most effective feature sets are presented in the table.

From the upper part of the table, the features including prosody from the preceding utterance (pLUU) are effective in the case of turn-keeping. This is because the speaker suggests turn-holding with the prosody of the previous utterances and that affects the filler generation. On the other hand, in the case of turn-switching, we see that the linguistic features (L) of the following utterance (fLUU) are indispensable. This suggests that the speaker generates a filler after deciding to take a turn, and the form of the filler is determined based on the DA and the approximate length of the following utterance.

Table 5 Feature set

Utterance	Type	Feature	
Preceding Utterance (pLUU)	Linguistic (L)	- DA	
		- POS of the last word	
		- Surface of the last word (if POS is an auxiliary verb or a post positional particle)	
		- Clause boundary	
		- # of words	
		- # of chunks	
	Prosodic (P)	F0	- Regression coefficient
			- Maximum value
		Power	- Minimum value
			- Mean value
		- Duration	
		- Speech rate	
		- Pause	
Following Utterance (fLUU)	Linguistic (L)	- DA	
		- POS of the beginning word	
		- Surface of the beginning word (if POS is a conjunction)	
		- # of words (quantized value)	
		- # of chunks (quantized value)	

The filler prediction model individually trained for each DA pair achieves better performance than the single model for all and assigns the typical form. The overall performance is not necessarily high, but this is due to arbitrary characteristics of fillers; fillers may be placed or not depending on the person and at different times.

4.4 Prediction in speech collision cases

Next, we conduct an experiment whether we can generate fillers to avoid speech collisions between two speakers using the model developed in the last subsection.

Speech collisions are defined as below.

- Both speeches are overlapped for over 500 ms after the beginning of the following utterance.
- The speaker who does not stop speaking after the speech collision takes the turn and the DA of his/her utterance is treated as the following DA.

There are 95 speech collisions in the relevant DA pairs in the corpus and the ratio of occurrence of fillers among these cases is 25.3% (=24/95). As the result of prediction, the proposed model can generate fillers in 57.9% (=55/95) of all cases in the speech collisions. This figure is around 2.3 times larger than the original number in the corpus, and covers 41.7% of them. Note that our model predicts fillers in 34.6% of DA pairs on average, close to the ratio of filler occurrence shown in Table 1.

Table 6 Prediction performance (F-measure)

DA pair	Feature set	Class	Individual	Single
RQ Keep	pLUU + fLUU LP + L	p	0.26	0.26
		o	0.32	0.00
		nf	0.50	**0.67**
		Avg.	**0.36**	0.31
RR Keep	pLUU + fLUU LP + L	d	**0.31**	0.25
		o	0.23	0.00
		nf	0.65	**0.74**
		Avg.	**0.40**	0.33
RS Keep	pLUU + fLUU L + L	d	0.17	0.17
		o	0.36	0.00
		nf	0.45	**0.69**
		Avg.	**0.33**	0.29
SS Keep	pLUU + fLUU LP + L	f	0.35	**0.36**
		nf	0.67	0.67
		Avg.	0.51	**0.52**
QR Switch	fLUU L	n	**0.33**	0.32
		o	0.50	0.00
		nf	0.75	**0.78**
		Avg.	**0.53**	0.37
RQ Switch	fLUU L	n	**0.53**	0.40
		o	0.41	0.00
		nf	0.37	**0.56**
		Avg.	**0.44**	0.32
RR Switch	pLUU + fLUU L + L	f	**0.59**	0.57
		nf	0.70	**0.71**
		Avg.	0.64	0.64
RS Switch	pLUU + fLUU LP + L	p	**0.20**	0.18
		o	0.44	0.00
		nf	**0.60**	0.58
		Avg.	**0.41**	0.25
SR Switch	fLUU L	f	**0.63**	0.50
		nf	**0.75**	0.70
		Avg.	**0.69**	0.60

p: proper form, d: demonstrative form, n: notice form
f: filler, o: other form, nf: no filler

The fillers generated by the proposed model can potentially avoid speech collisions. Even if the filler collides with the utterance by the dialog partner, it does not cause serious harm. This is particularly important for spoken dialog systems which cannot usually cancel the speech output command once generated.

5 Subjective evaluation

Finally, we conduct a subjective evaluation experiment on the fillers generated by the proposed method. We prepare ten audio samples with inserted fillers (with different frequencies) and no filler, respectively. These dialog segments are extracted from the corpus, and we had 20 people listen to the audio and make an evaluation on the questionnaire regarding to the naturalness and likability.

The result of multiple test shows that a significant difference is confirmed between the no-filler samples and the samples in which the frequency of fillers is low ($t(28)=4.62$ $p<.01$). In the case that the frequency of fillers is high, however, the difference is not significant ($t(28)=5.16$ $p<.1$). The result suggests that the user feels positive towards the system which generates fillers, but it might not be good to generate fillers too much.

6 Conclusions

We have proposed a prediction and generation mechanism of fillers in spoken dialog, which can suggest turn-holding or turn-taking. First, we found out that tendency of filler occurrence and its typical form is different according to the DA pairs. Based on this observation, we prepared a model to predict fillers for each DA pair, using both linguistic and prosodic features of the preceding utterance and the approximate linguistic features of the following utterance. It is shown that the effective feature set for prediction is different according to the DA pairs, in particular turn-holding or turn-taking. It is also shown that the DA features of the following utterance are useful, but this can be considered in the system design so that it determines the DA of the next utterance and then decide to generate a filler. Moreover, it is confirmed that the proposed model successfully generates fillers in more than half cases of speech collision, thus potentially avoid them or mitigate their side effect. We plan to implement this model in the spoken dialog system by an autonomous android ERICA [15, 14], and evaluate the effectiveness in real user experiences.

References

1. Akita, Y., Kawahara, T.: Statistical transformation of language and pronunciation models for spontaneous speech recognition. IEEE Transactions on Audio, Speech, and Language Processing **18**(6), 1539–1549 (2010)
2. Andersson, S., Georgila, K., Traum, D., Aylett, M., Clark, R.: Prediction and realisation of conversational characteristics by utilising spontaneous speech for unit selection. In: Proc. Speech Prosody (2010)
3. Bunt, H., Alexandersson, J., Carletta, J., Chae, J.W., Fang, A.C., Hasida, K., Lee, K., Petukhova, O., Popescu-Belis, A., Romary, L., et al.: Towards an iso standard for dialogue act annotation. in proceedings lrec 2010, malta (2010)

4. Cohen, J.: A coefficient of agreement for nominal scales. Educational and psychological measurement **20**(1), 37–46 (1960)
5. Den, Y.: Some phonological, syntactic, and cognitive factors behind phrase-final lengthening in spontaneous japanese: A corpus-based study. Laboratory Phonology **6**(3-4), 337–379 (2015)
6. Den, Y., Koiso, H., Maruyama, T., Maekawa, K., Takanashi, K., Enomoto, M., Yoshida, N.: Two-level annotation of utterance-units in Japanese dialogs: An empirically emerged scheme. In: LREC (2010)
7. Inoue, K., Milhorat, P., Lala, D., Zhao, T., Kawahara, T.: Talking with erica, an autonomous android. In: Proc. SIGdial Meeting Discourse & Dialogue, pp. 212–215 (2016)
8. Itagaki, H., Morise, M., Nisimura, R., Irino, T., Kawahara, H.: A bottom-up procedure to extract periodicity structure of voiced sounds and its application to represent and restoration of pathological voices. In: MAVEBA, pp. 115–118 (2009)
9. Kawahara, H., Masuda-Katsuse, I., De Cheveigne, A.: Restructuring speech representations using a pitch-adaptive time–frequency smoothing and an instantaneous-frequency-based f0 extraction: Possible role of a repetitive structure in sounds. Speech communication **27**(3), 187–207 (1999)
10. Kawahara, H., Morise, M., Takahashi, T., Nisimura, R., Irino, T., Banno, H.: Tandem-straight: A temporally stable power spectral representation for periodic signals and applications to interference-free spectrum, f0, and aperiodicity estimation. In: Acoustics, Speech and Signal Processing, 2008. ICASSP 2008. IEEE International Conference on, pp. 3933–3936. IEEE (2008)
11. Kawahara, T., Yamaguchi, T., Inoue, K., Takanashi, K., Ward, N.: Prediction and generation of backchannel form for attentive listening systems. In: Proc. INTERSPEECH, vol. 2016 (2016)
12. Koiso, H., Horiuchi, Y., Tutiya, S., Ichikawa, A., Den, Y.: An analysis of turn-taking and backchannels based on prosodic and syntactic features in japanese map task dialogs. Language and speech **41**(3-4), 295–321 (1998)
13. Koiso, H., Nishikawa, K., Mabuchi, Y.: Construction of The Corpus of Spontaneous Japanese (2006)
14. Lala, D., Milhorat, P., Inoue, K., Ishida, M., Takanashi, K., Kawahara, T.: Attentive listening system with backchanneling, response generation and flexible turn-taking. In: Proc. SIGdial Meeting Discourse & Dialogue, pp. 127–136 (2017)
15. Milhorat, P., Lala, D., Inoue, K., Tianyu, Z., Ishida, M., Takanashi, K., Nakamura, S., Kawahara, T.: A conversational dialogue manager for the humanoid robot ERICA. In: Proc. Int'l Workshop Spoken Dialogue Systems (IWSDS) (2017)
16. Ohsuga, T., Horiuchi, Y., Nishida, M., Ichikawa, A.: Prediction of turn-taking from prosody in spontaneous dialogue. Transactions of the Japanese Society for Artificial Intelligence **21**, 1–8 (2006)
17. Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., et al.: Scikit-learn: Machine learning in python. Journal of Machine Learning Research **12**, 2825–2830 (2011)
18. Schegloff, E.A., Sacks, H.: Opening up closings. Semiotica **8**(4), 289–327 (1973)
19. Shiwa, T., Kanda, T., Imai, M., Ishiguro, H., Hagita, N.: How quickly should communication robots respond? In: Human-Robot Interaction (HRI), 2008 3rd ACM/IEEE International Conference on, pp. 153–160. IEEE (2008)
20. Skantze, G., Hjalmarsson, A., Oertel, C.: Turn-taking, feedback and joint attention in situated human-robot interaction. Speech Communication **65**, 50–66 (2014)
21. Sundaram, S., Narayanan, S.: Spoken language synthesis: Experiments in synthesis of spontaneous monologues. In: In Proc. IEEE Workshop on Speech Synthesis, pp. 203–206 (2002)
22. Watanabe, M.: Features and Roles of Filled Pauses in Speech Communication: A corpus-based study of spontaneous speech. Hitsuji Syobo Publishing (2009)