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

Analysis of effect and timing of fillers in natural turn-taking

Divesh Lala, Shizuka Nakamura, Tatsuya Kawahara

Kyoto University, Graduate School of Informatics, Kyoto, Japan

{lala, shizuka, kawahara}@sap.ist.i.kyoto-u.ac.jp

Abstract

Turn-taking for spoken dialogue systems is still below the speed of real human conversation due to latency in speech and natural language processing, but fillers can be used by the system to take the turn more quickly without sacrificing naturalness. In this work we analyze fillers which are used at the start of turns in conversation and determine a window of appropriate times to use them. We analyze a human-robot conversation corpus to obtain an average response time of the fillers, and find that this differs according to the filler's form. We then conduct a subjective experiment in which participants dynamically change the timing of responses with and without fillers to designate a window of acceptable response timings. Our results show that the most suitable response time is around 200-500ms after the previous speaker has finished their turn. We also find differences in timing windows depending on existence of a filler used to begin the turn and its particular form. The implications of these results on the design of conversational systems are also discussed.
Index Terms: filler, turn-taking, human-computer interaction

1. Introduction

Fillers, or filled pauses, are a common occurrence in everyday conversation in a wide range of languages and serve various functions. Since their usage is so common, fillers should be one of the integral parts of conversational systems with a human embodiment, as opposed to systems such as smart speakers where fillers are not expected and are only redundant to answering questions from the user. Our long-term goal is to realize near-human conversational skill, which includes the use of fillers in a natural manner. This work is conducted with the human-like android robot ERICA [1].

Another problem addressed in this study is a smooth and natural turn-taking capability without an explicit push-to-talk or magic word interface. Fillers can assist with turn-taking regulation [2] but this has not been explored fully. A limited number of spoken dialogue systems use fillers to reduce overly long silences [3, 4, 5]. However in these systems the objective is to recover from a long period of silence, whereas fillers used for turn-taking should be used as a spontaneous but natural part of conversation. One research work used fillers in this manner, by waiting for a silence time of 500ms [6], but is still slow compared to real conversations.

Fillers can be used in a continuous turn-taking system [7, 8, 9] to initially take the turn while waiting for a speech recognition result. However, previous work does not indicate when it is appropriate to use fillers in this manner. We investigate this issue in this paper by focusing on the timing of turn-taking fillers, which we define as the filler used when starting a turn.

This work will analyze turn-taking fillers through a subjective experiment in which participants designate appropriate response times. In particular, we address the following research questions:

- What is the window of suitable response times for turn-taking fillers?
- Is there a difference in the window of response times depending on the form of turn-taking filler?
- Is there a difference in the window of response times depending on the existence of a turn-taking filler?

Answering these questions will help us to understand when a particular form of turn-taking filler can be said. This will be of interest for designers of conversational dialogue systems, since adding fillers with correct timing is necessary in spontaneous conversations [10] . Furthermore, research has shown that users are more positive towards a system which uses fillers than one which does not, since the system is seen as being more natural and has more social presence [11, 12, 13].

The relationship between fillers and turn-taking has been addressed in other works, which thoroughly analyzed acoustic features of fillers in a large corpus of English and Slovak [14, 15]. Other work has also focused on filler generation by using the previous dialogue act of the speaker [12]. In this work, we perform experiments with human subjects and explore how fillers affect their perception of turn-taking speed, with Japanese being the target language.

2. Analysis of data corpus

We first analyzed the fillers that are used in turn-taking and their timings, to get an understanding of these phenomena in human conversation.

2.1. Data collection

Our target data set is a collection of one-to-one conversations between a human subject and a tele-operated android ERICA. The operator of ERICA was one of a small number of trained actresses who were instructed to perform in a particular role. The types of conversations included attentive listening, speed dating and job interviews. Subjects ranged from students to elderly people, both male and female. Each session lasted between 5 and 15 minutes. We collected audio data and transcribed 85 sessions, including fillers and turn switches.

2.2. Analysis

We have previously analyzed the timing of turn-taking in the corpus [16] and found that the average silence time between turns is close to 100ms with many examples of overlapping turn switches, which is in line with other research on turn-taking speeds [17]. We defined a turn-taking filler as a filler which occurs at the beginning of a turn. Furthermore, we limited the analysis to turn-switches from the operator to subjects to reduce sample bias, since there were only a small number of operators.

http://dx.doi.org/10.21437/Interspeech.2019-1527

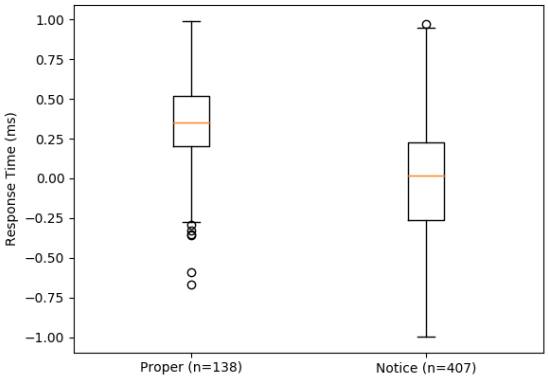

Figure 1: *Response times for proper and notice fillers. Times are calculated from when the previous speaker completely finishes their speech.*

In total, there were 4,174 subject turns, in which 816 (16.4%) contained turn-taking fillers. We used previous work as a guide to categorize the form of turn-taking fillers [12]. Using this categorization, we analyzed the distributions of filler forms in turn-taking fillers and non turn-taking fillers. There is a statistically significant difference in response times for the types of fillers used for turn-taking, as shown in Table 1.

Table 1: *Filler distribution occurrence ratio in corpus. TT means turn-taking fillers*

Filler form	Example (Japanese)	TT%	nonTT%
Proper	*um (etto)*	20.8	22.3
Demonstrative	*so (ano)*	4.4	29.5
Adverb	*well (ma-)*	5.6	21.6
Notice	*oh (a)*	53.8	8.6
Other	*wow (ee-)*	15.4	18.0
		100.0	100.0

Notice fillers are the most common for turn switching, but are less prominent for non-switching. The opposite is the case for demonstrative fillers. Adverb fillers are also more often used in non-switching situations. Proper fillers have approximately the same ratio in both turn switches and non-switches. Following on from this analysis, we decided to further analyze the timing of the most common forms of turn-taking fillers - notice and proper. We measured the time between the end of the operator's turn and the beginning of the subject's turn, for responses which began within one second before and after the previous speaker had finished. Altogether this totaled 712 turn-taking fillers. Figure 1 shows these results.

Notice turn-taking fillers are generally produced faster than proper turn-taking fillers, with the median time close to zero milliseconds. A Mann-Whitney test showed that there is a statistically significant difference between the medians of both distributions (p-value < 0.01), with an average difference in response time of 337ms. Both forms of turn-taking fillers are produced with little silence, if any, after the end of the previous speaker's turn and overlap is common.

From the perspective of system-based turn-taking, the latency from processing of voice activity detection adds delays in the reception of an automatic speech recognition result[1]. Other processing delays include generating a response from a database or a remote repository. This means that generating an answer at human-like speed is not feasible, and therefore the use of a filler to take the turn is a suitable solution.

We also found a difference in response times depending on the form of the turn-taking filler. These same forms will be tested in the subjective experiment.

3. Subjective evaluation experiment

We conducted an experiment to determine the appropriate timing for turn-taking fillers in a natural conversation. We designed a simple program which played back short audio samples of conversations contained in our corpus. These samples were of the operator (Speaker A) finishing their turn, and then the interlocutor (Speaker B) taking the turn.

Four categories of samples were used for this experiment. The first category (**PROPER**) had Speaker B using a proper filler to take the turn. The second category (**NOTICE**) had Speaker B using a notice filler to take the turn. The other two categories (**PROPER-NF** and **NOTICE-NF**) were the same samples as the first two categories, except that the corresponding fillers were removed from Speaker B's response. The number of potential samples was eight for each category.

The choice of samples is important in this experiment. We chose samples in which turn-taking was "smooth". That is, timing of Speaker B's response is similar to the typical response time we found in our corpus, did not begin with consecutive fillers or other disfluencies. Samples were short segments of conversation (around 10 seconds) since the subjects would have to listen to them many times. Furthermore, we only used samples in which Speaker B used the appropriate filler according to Speaker A's dialogue act. As discussed in Section 2.2, the use of proper and notice fillers after a question were suitable, but only questions which are not easily answerable, such as "What is your name?".

An example of a sample dialogue from the **PROPER** category is:

A: "At university, do you have a major and research topic?"

B: "Um, I'm in the agriculture department..."

and from the **NOTICE** category:

A: "Can you sing those songs?"

B: "Oh, it's impossible for me."

Corresponding **PROPER-NF** and **NOTICE-NF** samples removed the "um" and "oh". Subjects evaluated 12 samples in total, with four samples being randomly selected from each of the **PROPER** and **NOTICE** categories and two from each of the **PROPER-NF** and **NOTICE-NF** categories. It was ensured that the same subject would not evaluate a non-filler sample and its **PROPER** or **NOTICE** version.

A screenshot of the experiment program is shown in Figure 2, translated to English. The slider allows the subject to manipulate the timing of Speaker B's response. When the "Play" button is pressed, Speaker A will begin their utterance and Speaker B's response would begin according to the slider's value. The "Stop" button could be pushed to stop the sample so that the subjects did not have to repeatedly listen to the whole recording. The timing is adjusted in 100ms increments, although this

[1]The system cannot detect the end of a user utterance before detection of a pause.

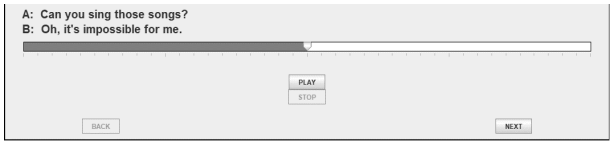

Figure 2: *GUI of the program used in the experiment (translated from Japanese). The slider can by manipulated by the subject to change the response time of Speaker B.*

was not explicitly shown. The limits of Speaker B's utterance timing was ± 2 seconds relative to the end of Speaker A's utterance.

For every sample, the subjects performed three tasks. The first task was to move the slider so that in their opinion the timing of Speaker B's response was the earliest possible without sounding unnatural. The second task was to do the same for the latest possible timing. The third task was to move the slider to where they thought the timing was the most suitable. We will refer to these three response timing categories as **FAST**, **SLOW** and **BEST**. 31 subjects participated in the experiment (18 male, 13 female), the majority being university students.

There are several limitations to this methodology, the most critical being that the samples which were evaluated have no context. A longer segment of conversation with multiple turns is more suitable, but this takes much more time and is prone to subject fatigue. We decided that a simpler experiment would be easier for subjects to understand. Furthermore, there are many more factors that influence response times, including the prosody and speech rate of the previous utterance as well as subject demographics. Analyzing each of these factors individually requires a more comprehensive experiment and greater number of samples.

4. Results

We present our results based on our research questions posed in Section 1.

4.1. Window of turn-taking response times

We first conduct a basic analysis of the times that subjects determined as being **FAST**, **SLOW** and **BEST**. The results are shown using boxplots in Figure 3. Boxes represent the interquartile range. The whiskers represent responses which are within $Q1 - (1.5 * IQR)$ and $Q3 + (1.5 * IQR)$, with $Q1$ and $Q3$ the 1st and 3rd quartiles and IQR the interquartile range.

The window of **BEST** times is in general between the **FAST** and **SLOW** times. These results suggest that a suitable window of response times for fillers is 200-500ms and is slightly slower than the actual response times found in the corpus. An "acceptable" window of response times (**FAST** median to **SLOW** median) is 400ms before to 900ms after the previous speaker has stopped. There is still much variability in the data, both due to individual subjectivity and the samples themselves.

We also analyzed the individual standard deviations of every subject's response times and compared these to the individual standard deviations of every filler sample. The results are shown in Figure 4.

Subjects showed less variability in their responses than the variability for each sample. This suggests that individuals tended to select similar windows for all samples, but these windows can be quite different from each other for a given sample.

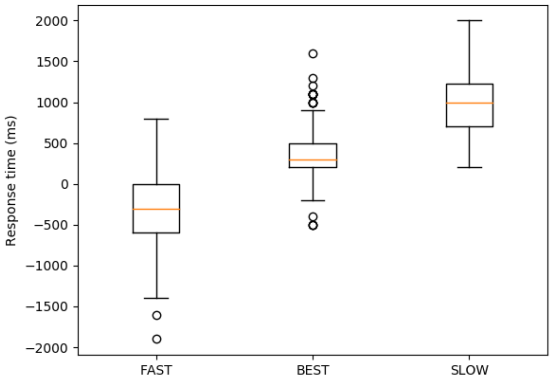

Figure 3: *Subjectively evaluated times for each timing response category.*

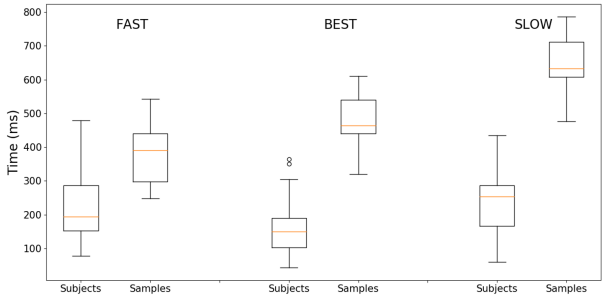

Figure 4: *Standard deviations of times for subjects and samples.*

There is large variation in the acceptable range of a particular sample's **SLOW** timing response.

4.2. Differences in filler forms

In Table 2 we compare the **PROPER** and **NOTICE** samples over the three timing categories. Shapiro-Wilk tests indicated non-normality, so Mann-Whitney tests were used for statistical comparisons. Median and interquartile range are reported.

We find that there is a statistically significant difference between the two filler categories for **BEST** and **SLOW** timing, but not for **FAST**. This result indicates that the window in which a notice filler is deemed acceptable is smaller and finishes earlier. Proper fillers can be spoken almost 300ms later than notice fillers while still having acceptable timing.

4.3. Differences in filler and no-filler timing

Similar to the analysis in the previous section, we compare all filler samples to all no-filler samples, categorized by the filler form. Results are shown in Table 3.

We find that there was differences in all comparisons, except the **PROPER** and **PROPER-NF** samples for the **SLOW** response time. The largest differences were in the **FAST** response time, where removing a filler delayed the start of the timing window as much as 300ms on average. This indicates fillers can be overlapped with the preceding user's utterances, but the utterances themselves cannot be.

On the other hand, there was little difference between fillers and no-filler samples in the **SLOW** response category. This suggests that the end of the timing window is not affected by

Table 2: *Comparison of filler forms for each timing category.*

	FAST	BEST	SLOW
PROPER			
Median	-400	350	1100
Interquartile range	600	300	600
NOTICE			
Median	-500	200	800
Interquartile range	600	225	400
Mann-Whitney p-value	0.069	<**0.001***	<**0.001***

Table 3: *Comparison of filler and no-filler samples for each timing category. Mann-Whitney p-values relate to the corresponding category which included fillers.*

	FAST	BEST	SLOW
PROPER-NF			
Median	-200	400	1100
Interquartile range	500	300	400
Med. diff. from **PROPER**	+200	+50	0
Mann-Whitney p-value	**0.009***	**0.018***	0.247
NOTICE-NF			
Median	-200	300	900
Interquartile range	600	300	400
Med. diff. from **NOTICE**	+300	+100	+100
Mann-Whitney p-value	<**0.001***	**0.001***	**0.001***

the existence of fillers. This means a filler can be used to buy time for long latencies in response generation.

We summarize our results in Figure 5. The most suitable timing window is determined as the approximate central range of all the **BEST** response times.

5. Analysis and discussion

This study provided a window of response times in which we can use fillers for turn-taking. We found that a response time of 200-500ms was suitable for both turn-taking with fillers and without. Although our results showed a large window of acceptable response times, the median of the slowest acceptable timing regardless of fillers is around one second. To mitigate the problem of longer delays, a filler can be used before the turn is taken without loss of naturalness. This finding is important for systems which require time to generate responses, such as retrieval from an online database.

We also found that adding a filler increases the length of the timing response window by extending the beginning of it to an earlier time point. This suggests that fillers are not considered to be as intrusive as using a no-filler response, and so can be used earlier. This confirms the functional value of turn-taking fillers as an option to take the turn quickly while preparing the actual response.

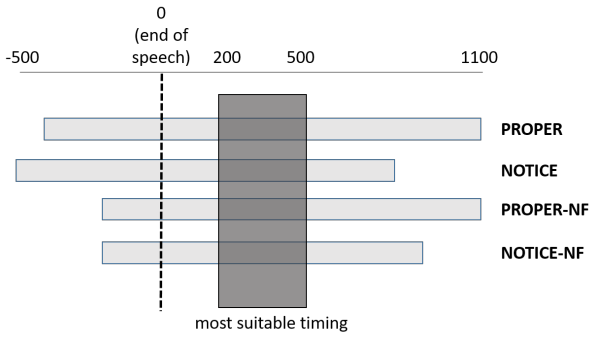

Figure 5: *Average acceptable timing windows for each sample category. A most suitable timing window is added for reference.*

It is shown that the window of response times is later for proper turn-taking fillers than notice turn-taking fillers, as reflected in the original corpus. The reason for this may be due to the differences in the semantic meaning of the fillers. Replacement with another filler form could affect the timing response. The choice of filler is greatly dependent on the utterance itself, and requires further studies.

Our goal is to develop a turn-taking model for an android robot, and for human-like turn-taking behavior we should aim to take the turn in the window of suitable response times. A continuous model, which predicts the end of turn while the user is speaking, is appropriate for this because turn switching occurs within a short period of silence after the previous turn or even before the previous turn has ended. With a continuous model, fillers can be used to take the turn without needing to wait for response generation. A robust model of this kind would exhibit human-like turn-taking speed.

This work ignored non-verbal modalities such as gaze which are also used to coordinate turn-taking, but it is clear that both are needed for a system to come close to human conversation. The timing of these non-verbal behaviors with fillers should also be kept in mind since synchronization with speech is critical.

6. Conclusion

In this work we presented an analysis of suitable timing responses for turn-taking fillers. We found that a window of suitable timing was around 200-500ms for both filler and non-filler responses, and also found differences in timing windows depending on the form and existence of a filler. We presented implications for the design of conversational robots and propose that using fillers as a means to take the turn is natural and can be done to buy time while the system prepares a response, even after a relatively long silence between turns. The results of this work will be used to assist in the implementation of an online turn-taking model for a conversational robot.

7. Acknowledgements

This work was supported by JST ERATO Ishiguro Symbiotic Human-Robot Interaction program (Grant Number JPMJER1401), Japan.

8. References

[1] D. F. Glas, T. Minato, C. T. Ishi, T. Kawahara, and H. Ishiguro, "Erica: The erato intelligent conversational android," in *2016 25th IEEE International Symposium on Robot and Human Interactive Communication (RO-MAN)*. IEEE, 2016, pp. 22–29.

[2] 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.

[3] N. Mukawa, H. Sasaki, and A. Kimura, "How do verbal/bodily fillers ease embarrassing situations during silences in conversations?" in *The 23rd IEEE International Symposium on Robot and Human Interactive Communication*. IEEE, 2014, pp. 30–35.

[4] N. Ohshima, K. Kimijima, J. Yamato, and N. Mukawa, "A conversational robot with vocal and bodily fillers for recovering from awkward silence at turn-takings," in *2015 24th IEEE International Symposium on Robot and Human Interactive Communication (RO-MAN)*. IEEE, 2015, pp. 325–330.

[5] T. Shiwa, T. Kanda, M. Imai, H. Ishiguro, and N. Hagita, "How quickly should a communication robot respond? delaying strategies and habituation effects," *International Journal of Social Robotics*, vol. 1, no. 2, pp. 141–155, 2009.

[6] G. Skantze, M. Johansson, and J. Beskow, "Exploring turn-taking cues in multi-party human-robot discussions about objects," in *Proceedings of the 2015 ACM on International Conference on Multimodal Interaction*. ACM, 2015, pp. 67–74.

[7] M. Roddy, G. Skantze, and N. Harte, "Multimodal continuous turn-taking prediction using multiscale rnns," in *Proceedings of the 2018 on International Conference on Multimodal Interaction*. ACM, 2018, pp. 186–190.

[8] K. Hara, K. Inoue, K. Takanashi, and T. Kawahara, "Prediction of turn-taking using multitask learning with prediction of backchannels and fillers," *Proc. Interspeech 2018*, pp. 991–995, 2018.

[9] G. Skantze, "Towards a general, continuous model of turn-taking in spoken dialogue using lstm recurrent neural networks," in *Proceedings of the 18th Annual SIGdial Meeting on Discourse and Dialogue*, 2017, pp. 220–230.

[10] M. Taboada, "Spontaneous and non-spontaneous turn-taking," *Pragmatics. Quarterly Publication of the International Pragmatics Association (IPrA)*, vol. 16, no. 2, pp. 329–360, 2006.

[11] H. Goble and C. Edwards, "A robot that communicates with vocal fillers has… uhhh… greater social presence," *Communication Research Reports*, vol. 35, no. 3, pp. 256–260, 2018.

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

[13] K. Ohta, N. Kitaoka, and S. Nakagawa, "Modeling filled pauses and silences for responses of a spoken dialogue system," *International Journal of Computers*, vol. 8, pp. 136–142, 2014.

[14] Š. Beňuš, "Variability and stability in collaborative dialogues: turn-taking and filled pauses," in *Tenth Annual Conference of the International Speech Communication Association*, 2009.

[15] Š. Benuš, "Cognitive aspects of communicating information with conversational fillers in slovak," in *2013 IEEE 4th International Conference on Cognitive Infocommunications (CogInfoCom)*. IEEE, 2013, pp. 271–276.

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

[17] T. Stivers, N. J. Enfield, P. Brown, C. Englert, M. Hayashi, T. Heinemann, G. Hoymann, F. Rossano, J. P. De Ruiter, K.-E. Yoon *et al.*, "Universals and cultural variation in turn-taking in conversation," *Proceedings of the National Academy of Sciences*, vol. 106, no. 26, pp. 10 587–10 592, 2009.