INTERSPEECH 2017
August 20–24, 2017, Stockholm, Sweden

Analysis of the Relationship between Prosodic Features of Fillers and Its Forms or Occurrence Positions

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

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

Abstract

Fillers are involved in the ease of understanding by listeners and turn-taking. However, the knowledge about its prosodic features is insufficient, and its modeling has not been done either. For these reasons, there is insufficient knowledge to generate natural and appropriate fillers in a dialog system at present. Therefore, for the purpose of clarifying the prosodic features of fillers, its relationship with occurrence positions or forms were analyzed in this research. 'Ano' and 'Eto' were used as forms, non-/boundary of Dialog Act and non-/turn-taking for occurrence positions. Duration, F0, and intensity were utilized as prosodic features. As a result, the followings were found out: the prosodic features are different depending on the difference of the occurrence positions even for fillers of the same form, and similar prosodic features are found between the same occurrence positions even in different forms.

Index Terms: fillers, prosodic features, forms, occurrence positions, dialog

1. Introduction

The final purpose of this research is to generate natural and appropriate fillers in dialog by spoken dialog systems or autonomous androids. For such generation, it is necessary to correctly grasp the following characteristics of fillers: occurrence positions, forms, and prosody.

In the researches [1-3] on the functions of each form of fillers, the relationship between the functions of fillers and the syntactic/semantic characteristics of the utterances succeeding the fillers is handled. Moreover, in the study [4] analyzing the relationship between the prosodic features of fillers and the characteristics of the accent phrases of the utterances preceding and succeeding the fillers, not only succeeding ones but also preceding ones are handled. Furthermore, in the analysis of fillers, it is also important to consider not only the relation with the utterances preceding and succeeding the fillers but also the traits of utterance boundaries where fillers occur.

However, these studies mainly focus on monolog. Unlike in monolog, speakers are changed in dialog. Therefore, it is also necessary to consider the difference of occurrence positions in dialog based on turn-taking/holding. Regarding such occurrence positions of fillers, turn-taking/holding in addition to the relation with the utterances preceding and succeeding the fillers have been considered in some studies [5]. In the study, in order to grasp the tendency of occurrence positions of fillers, Dialog Act (hereafter, DA) is used to represent an occurrence position, and it is clarified that representative filler forms are different according to the type of DA sequences.

However, in order to generate fillers, it is necessary that not only selecting appropriate forms but also controlling prosodic features of fillers themselves. Therefore, analyses in this research are stressed on prosodic features of fillers. Then, in order to make the comparison of prosodic features clearer, typical forms in Japanese language such as 'Ano' [ano] and 'Eto' [eto] among various forms are focused on.

2. Dialog corpus

Dialogs by two people who met for the first time each other are used for analyses in this research. An example of the recording environment is shown in Figure 1. These dialogs were chat simulated by a speaker as the role of a secretary of a university laboratory (right) and the other as the visitor of the laboratory (left) [5]. Regarding the role of the secretary, the operator remotely controlled the android ERICA [6] facing the visitor from a separate room, and the voice of the operator was reproduced as it was for the voice of the android.

The voice of the operator is analyzed in this research. For the purpose of clearly grasping the differences in prosodic features within the same speaker, the speech by one speaker (female, 30s) of the Tokyo dialect is targeted. The data length is about 10 minutes per dialog, about 50 minutes of five dialogs in total. The different person played the role of the visitor in each dialog.

In case of turn-taking, utterances between speakers are easy to overlap. If the ending time of the filler by the operator is later than the ending time of the utterance by the speaker as the visitor, the filler by the operator is a target to be analyzed though they are overlapped.

Figure 1: *An example of the recording environment of a dialog between a speaker as the role of a secretary of a university laboratory (right) and the other as the visitor of the laboratory (left).*

http://dx.doi.org/10.21437/Interspeech.2017-1006

Table 1: *Occurrence frequency and ratio in each form in about 50 minutes of five dialogs.*

Form written in Japanese 'Hiragana' and [IPA]	Occurrence Frequency	Ratio [%]	
		Including 'あ' [a]	Not including 'あ' [a]
'あの(ー)' [ano(:)]	195	50.4	63.7
'あ' [a]	81	20.9	-
'え(ー)(っ)と(ー)' [e(:)(t)to(:)]	66	17.1	21.6
'なんか' [naŋka]	14	3.6	4.6
Others	31	8.0	10.1
Total	387	100.0	100.0

Table 2: *Occurrence frequency and ratio of each form at the occurrence position of the DA boundary and the DA non-boundary.*

DA boundary

Form written in Japanese 'Hiragana' and [IPA]	Occurrence Frequency	Ratio [%]	
		Including 'あ' [a]	Not including 'あ' [a]
'あ' [a]	77	49.4	-
'あの(ー)' [ano(:)]	37	23.7	46.8
'え(ー)(っ)と(ー)' [e(:)(t)to(:)]	23	14.7	29.1
Others	19	12.2	24.1
Total	156	100.0	100.0

DA non-boundary

Form written in Japanese 'Hiragana' and [IPA]	Occurrence Frequency	Ratio [%]	
		Including 'あ' [a]	Not including 'あ' [a]
'あの(ー)' [ano(:)]	158	68.4	69.6
'え(ー)(っ)と(ー)' [e(:)(t)to(:)]	43	18.6	18.9
'なんか' [naŋka]	14	6.1	6.2
'あ' [a]	4	1.7	-
Others	12	5.2	5.3
Total	231	100.0	100.0

3. Analysis of forms and occurrence positions

3.1. Characteristics of forms

In Table 1, all the fillers to be analyzed are arranged in descending order of occurrence frequency. The category of fillers defined taking into consideration forms and functions contains the followings: Peculiar (e. g., 'えーと' [e:to], 'えと' [eto]), Demonstrative (e. g., 'あのー' [ano:], 'そのー' [sono:]), Adverb (e. g., 'まー' [ma:], 'なんか' [naŋka]), Awareness (e. g., 'あ' [a], 'え' [e]), and Other (e. g., 'なんていうか' [nantejɯ:ka]) [5]. As shown in Table 1, the following categories occupy over 80% of the total: Demonstrative, Awareness, and Peculiar.

Forms with high occurrence frequency in each category are analyzed in this research. However, 'あ' (hereafter, 'A') [a] is excluded from the analysis of prosodic features in this study, because it might be regarded as not a typical example of filler but a kind of a response word, although it is frequently occurred. Then, the right-hand side of Table 1 shows the ratios when not including 'A.'

In addition, similar forms in each category are handled together. Specifically, fillers of the form 'あの(ー)' [ano(:)] in a Demonstrative category is classified into the group 'Ano' (hereafter, G-Ano), and fillers of the form 'え(ー)(っ)と(ー)' [e(:)(t)to(:)] in a Peculiar category is classified into the group 'Eto' (hereafter, G-Eto). Fillers classified into these groups are analyzed in this research.

3.2. Annotation of occurrence positions

In order to identify the occurrence positions of fillers, annotation of DA was added to all of the speech with reference to the previous study [5]. One DA tag was given to each Long Utterance Unit [7], which is a syntactic, conversational, and interactive unit defined based on the Clause Unit [8].

The occurrence positions are divided into the DA boundary (hereafter, DA-B) and the DA non-boundary (hereafter, DA-nB) depending on whether it is a DA boundary or not. Furthermore, the DA-Bs are subdivided into the turn-taking position (hereafter, T-T) and the turn-holding position (hereafter, T-H) depending on whether the speaker changes or not. The total number of the DA-Bs was 606, including the positions where no filler occurred. In case of including 'A,' the occurrence rate was 25.7% since the occurrence frequency was 156. In case of not including 'A,' the occurrence rate was 13.0% since the occurrence frequency was 79.

3.3. Occurrence tendency of each form by occurrence position

In Table 2, the fillers occurred at the DA-B and the DA-nB are arranged in descending order of occurrence frequency. In case of not including 'A,' the G-Ano is 46.8%, and the G-Eto is 29.1%, at the DA-B. On the other hand, the G-Ano is 69.6%, and the G-Eto is 18.9%, at the DA-nB. In other words, it finds out that the G-Ano tends to occur more than the G-Eto, at the DA-nB compared with the DA-B.

In Table 3, the fillers occurred at the T-T and T-H are arranged in descending order of occurrence frequency. In short, Table 3 shows the breakdown of the DA-B in Table 2. In case of not including 'A,' the G-Ano is 50.0%, and the G-Eto is 18.2%, at the T-T. On the other hand, the G-Ano is 45.6%, and the G-Eto is 33.3%, at the T-H. In other words, it finds out that the G-Ano tends to occur more than the G-Eto, at the T-T compared with the T-H.

Table 3: *Occurrence frequency and ratio of each form at the occurrence position of the turn-taking and the turn-holding.*

Turn-taking

Form written in Japanese 'Hiragana' and [IPA]	Occurrence Frequency	Ratio [%]	
		Including 'あ' [a]	Not including 'あ' [a]
'あ' [a]	65	74.7	-
'あの(ー)' [ano(:)]	11	12.6	50.0
'え(ー)(っ)と(ー)' [e(:)(t)to(:)]	4	4.6	18.2
Others	7	8.0	31.8
Total	87	100.0	100.0

Turn-holding

Form written in Japanese 'Hiragana' and [IPA]	Occurrence Frequency	Ratio [%]	
		Including 'あ' [a]	Not including 'あ' [a]
'あの(ー)' [ano(:)]	26	37.7	45.6
'え(ー)(っ)と(ー)' [e(:)(t)to(:)]'	19	27.5	33.3
'あ' [a]	12	17.4	-
Others	12	17.4	21.1
Total	69	100.0	100.0

Taken together, there is a possibility of being two types for the G-Ano. Figure 2 shows the relationship between two types of the G-Ano and the occurrence positions. One is the 'Type 1' at the T-T, and the other is 'Type 2' at the DA-nB. This point will be discussed again in 4.2.

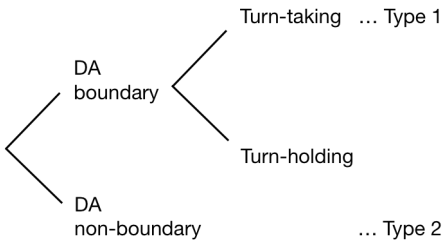

Figure 2: *A possibility of being two types in the group 'Ano.'*

4. Analysis of prosodic features

4.1. Measuring methods

4.1.1. Duration

The duration of each filler is calculated from the starting and the ending time of the relevant filler.

4.1.2. Fundamental Frequency

The fundamental frequency (hereafter, F0) of each filler is calculated using TANDEM-STRAIGHT [9] (XSX [10]). In consideration of characteristics of human beings, its common logarithm is used for analyses. As an approximate value representing the height and its change, the following values of F0 are used: averages and ranges.

The range of the F0 $Range_{F0,n}$ of a certain sample filler n is expressed by the following equation:

$$Range_{F0,n} = F0_{max,n} - F0_{min,n} .$$

Here, $F0_{max,n}$ is the maximum of the F0 of a certain sample filler n, and $F0_{min,n}$ is the minimum of the F0 of a certain sample filler n.

4.1.3. Intensity

The intensity of each filler is also calculated using TANDEM-STRAIGHT (XSX). The values outputted with the default settings are used for analyses. In consideration of characteristics of human beings, its maximum is used for analyses.

4.2. Comparison between occurrence positions in the same forms

Prosodic features of fillers in the same forms were compared between occurrence positions. T-test was conducted to investigate whether there is a significant difference between average values of the distribution of each prosodic feature.

In the G-Eto, there was no significant difference in prosodic features between occurrence positions. In other words, in the G-Eto, a kind of systematic bias of prosodic features depending on an occurrence position has not been observed at least at present moment. On the other hand, in the G-Ano, there was a significant difference in averages or ranges of F0 between some occurrence positions. Meanwhile, there was no significant difference in durations or intensities between occurrence positions. In other words, it was found that the difference of prosodic features by an occurrence position depends on a form. The results are shown in Table 4. From this table, the followings can be found:

- DA-B and DA-nB
  - (1) Averages of F0: DA-B > DA-nB
  - (2) Ranges of F0: DA-B < DA-nB
- T-T and T-H
  - (3) Averages of F0: T-T > T-H
- T-T and DA-nB
  - (4) Averages of F0: T-T > DA-nB
  - (5) Ranges of F0: T-T < DA-nB
- T-H and DA-nB
  - (6) Averages of F0: T-H > DA-nB .

Based on these findings, when comparing the averages or ranges of F0 between occurrence positions, the prosodic features of the G-Ano are as shown in Figure 3. Since the magnitude of each prosodic feature is shown on the vertical axis, the magnitude relation of them in each occurrence position can be compared. This figure suggests that there exist

Table 4: *Comparison of features of F0 in the group 'Ano' between occurrence positions. Only those with a significant difference.*

F0	(1) Average		(2) Range		(3) Average		(4) Average		(5) Range		(6) Average	
Occurrence position	DA-B	DA-nB	DA-B	DA-nB	T-T	T-H	T-T	DA-nB	T-T	DA-nB	T-H	DA-nB
# of samples	37	158	37	158	11	26	11	158	11	159	26	158
Average	2.31	2.26	0.14	0.20	2.35	2.30	2.35	2.26	0.10	0.20	2.30	2.26
Standard deviation	0.06	0.11	0.14	0.16	0.03	0.06	0.03	0.11	0.05	0.16	0.06	0.11
Significant difference	$p < 0.001$		$p < 0.05$		$p < 0.01$		$p < 0.001$		$p < 0.001$		$p < 0.01$	

Symbols
DA-B: DA boundary, DA-nB: DA non-boundary, T-T: Turn-taking, and T-H: Turn-holding

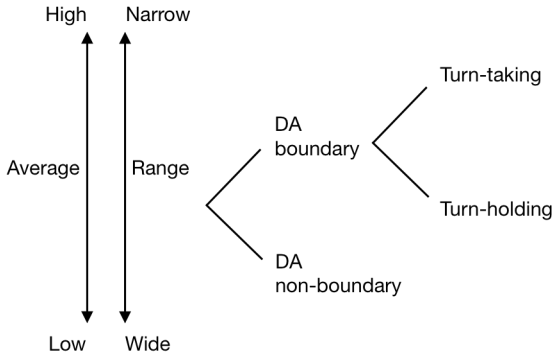

Figure 3: *Comparison of features of F0 in the group 'Ano'.*

two kinds of typical patterns of prosodic features for the G-Ano depending on an occurrence position. In other words, one shows the lower averages of F0 and the wider ranges of F0 at the DA-nB. The other shows the higher averages of F0 and the narrower ranges of F0 at the T-T. This agrees in possibility of being two types at the DA-nB and the T-T in the G-Ano which is shown in Figure 2 in 3.3. However, detailed analysis is necessary for the obtained result of the combination of the higher averages of F0 and the narrower ranges of F0 in the future.

4.3. Comparison between forms in the same occurrence positions

Prosodic features of fillers in the same occurrence positions were compared between forms. T-test was conducted to investigate whether there is a significant difference between average values of the distribution of each prosodic feature.

There was a significant difference at every occurrence position between durations. Further, there was a significant difference at the DA-B or T-H between the ranges of F0. The difference of duration is affected by the difference of a phoneme sequence based on the difference of a form. This difference also seems to affect the difference in the range of F0.

On the other hand, there was no significant difference in the average of F0 or intensity at any occurrence position between forms. In other words, similar prosodic features are sometimes found between the same occurrence positions even in different forms. This is also considered to be effective knowledge in controlling prosodic features of fillers.

5. Future works

In this research, though DA was used to represent an occurrence position, a category of DA itself was not considered. However, to grasp the functions and the characteristics of fillers in the flow of dialog, it has to be considered like in the previous study [5]. In addition, as mentioned above, the followings were analyzed for the prosodic features of fillers in monolog in the previous study [4]: the relationship between the prosodic features of fillers and the characteristics of the accent phrases of the utterances preceding and succeeding the fillers, and the duration between the filler and its preceding or succeeding utterances. Therefore, it is necessary to analyze them in dialog as well.

6. Acknowledgements

This study was supported by ERATO Ishiguro Symbiotic Human-Robot Interaction Project.

7. References

[1] T. Sadanobu and Y. Takubo, "The monitoring devices of mental operations in discourse--A case of 'eeto' and 'ano (o)'," *Journal of the Lnguistic Society of Japan*, vol. 108, pp. 74–92, 1995.

[2] M. Watanabe, *Features and Roles of Filled Pauses in Speech Communication: A corpus-based study of spontaneous speech*. Tokyo: Hitsuji Shybo, 2009.

[3] T. Kawada, *On the speech form of Japanese fillers and its characteristics – Using the degree of interaction with listeners as an indicator*. Ph. D. thesis, Kyoto University, DOI: 10.14989/doctor.k15563, 2010.

[4] K. Maekawa, "Preliminary study on the characteristics of filled pauses in spontaneous speech: Analysis of location and pitch height," *Proceedings of the Phonetic Society of Japan*, vol. 16, no. 3, pp. 106–107, 2012.

[5] R. Nakanishi, K. Inoue, S. Nakamura, K. Takanashi, and T. Kawahara, "Predicting occurrence and form of fillers based on Dialog Act pairs for smooth turn-taking," *Proceedings of the Japanese Society for Artificial Intelligence*, SIG-SLUD-B506-04, pp. 18-24, 2017.

[6] K. Inoue, P. Milhorat, D. Lala, T. Zhao and T. Kawahara, "Talking with ERICA, an autonomous android," *Proceedings of the SIGdial Meeting on Discourse and Dialogue*, pp. 212-215, 2016.

[7] The Japanese Discourse Research Initiative, *Utterance-Unit Labeling Manual version 2.0*. http://www.jdri.org/resources/manuals/uu-doc-2.0.pdf, 2014.

[8] T. Maruyama, K. Takanashi, and K. Uchimoto, Clause Unit Information. in National Institute for Japanese Language and Linguistics (Eds.) *Construction of The Corpus of Spontaneous Japanese*, http://pj.ninjal.ac.jp/corpus_center/csj/k-report-f/CSJ_rep.pdf, pp. 255-321, 2006.

[9] H. Kawahara, M. Morise, T. Takahashi, R. Nisimura, T. Irino, and H. Banno, "TANDEM-STRAIGHT: A temporally stable power spectral representation for periodic signals and applications to interference-free spectrum, F0, and aperiodicity estimation," *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing*, pp. 3933–3936, 2008.

[10] H. Itagaki, M. Morise, R. Nisimura, T. Irino, and H. Kawahara, "A bottom-up procedure to extract periodicity structure of voiced sounds and its application to represent and restoration of pathological voices," *Proceedings of the International Workshop MAVEBA*, pp. 115–118, 2009.