10th International Conference on Speech Prosody 2020
25-28 May 2020, Tokyo, Japan

Analysis and modeling of between-sentence pauses in news speech by Japanese newscasters

Shizuka Nakamura[1], *Carlos Toshinori Ishi*[2], *Tatsuya Kawahara*[1]

[1]Graduate School of Informatics, Kyoto University, Kyoto, Japan
[2]Hiroshi Ishiguro Laboratories, ATR, Kyoto, Japan
shizuka@sap.ist.i.kyoto-u.ac.jp, carlos@atr.jp, kawahara@i.kyoto-u.ac.jp

Abstract

Many speech synthesizers hardly consider between-sentence pauses. This could be one of the factors of the monotony of continuous synthesized speech. Aiming at breaking the monotony and improving the news speech likeness, we analyzed the characteristics of between-sentence pause durations of news speech by two newscasters and constructed a model to predict these durations. Analysis of the pause durations firstly revealed that the difference in the distributions between the two newscasters are largely affected by pauses after lead sentences, which have a large freedom. Then, from prosodic context analysis, it became clear that the following prosodic features have a correlation with between-sentence pause durations: the mean F0 of the last part in the preceding sentence, and the number of morae included in the subsequent sentence. The correlation coefficient between the predicted values by a linear multiple regression model using these parameters and the measured values was 0.44 for the test data. It was found that between-sentence pause durations could be predicted to some extent by utilizing prosodic information of the preceding and subsequent speech features. The news speech likeness of continuous synthesized speech can be improved by incorporating this model into existing speech synthesizers which generate speech sentence by sentence.

Index Terms: between-sentence pause, newscaster speech, Japanese

1. Introduction

Silent pauses in speech are not only physiologically or linguistically essential to speakers, but also indispensable to listeners to understand the speech contents [1-4]. In particular, in the case of news speech read by newscasters, pauses are considered to play an important role since easy understanding by listeners is required. Actually, in an interview by the authors, professional Japanese newscasters told that reading news manuscripts is a special style that is completely different from daily conversation. Details of the interview will be introduced in Section 2.2.

Previous studies on pauses have shown the existence of language dependency [5] and speaker dependency [1-2, 6-7]. Not only comparison of spontaneous and read speech [8-9], but also various speaking styles [10] have been conducted.

Several quantitative studies on pauses in Japanese have been carried out so far. For example, it has been reported that the duration of pauses is proportional to the duration of speech between exhalations [6]. The relationship between local phrase structures and pause insertion was analyzed in [7].

Based on an occurrence position, pauses in speech are divided into the following two types: i) between-sentence pauses, and ii) within-sentence pauses. Comparison of between-sentence and within-sentence pauses has been carried out in previous studies on Japanese speech. It is reported in [11] that the difference in the mean durations of between-sentence and within-sentence pauses is significantly large regardless of the speech rate. Although there has been little quantitative analysis on pauses in Japanese news speech, it is reported that the mean durations of within-sentence and between-sentence pauses in actual news speech are approximately 400 and 1,600 msec, respectively [12]. In addition, although not news speech of professional newscasters but normal reading speech, it is reported in [7] that within-sentence pause durations have positive correlation with its preceding durations of the speech. Moreover, it is shown that between-sentence pause durations are lengthened when a topic changes, and it is pointed out that the lengthening can be influenced by the physiological control such as breathing and swallowing.

In recent years, speech synthesis technology has made great strides, and the naturalness of synthesized speech has also been improved. On one hand, consideration for within-sentence pauses is often given since almost all of the speech synthesizers input and output sentence by sentence. On the other hand, consideration for between-sentence pauses is hardly given; for example, a duration of a between-sentence pause is often fixed in any case. Since this is considered one of the factors of the monotony of continuous synthesized speech, it is required to clarify the characteristics of between-sentence pauses and describe them mathematically. Few studies have systematically attempted to analyze and mathematically describe between-sentence pauses in Japanese news speech.

Based on these backgrounds and problems, aiming at breaking the monotony and improving the news speech likeness, we analyze the characteristics of between-sentence pause durations of news speech and construct a model to predict these durations. This model predicts the duration of between-sentence pauses using prosodic information of the preceding and subsequent speech. We expect that the news speech likeness of continuous synthesized speech is improved by incorporating this model into existing speech synthesizers which generate speech sentence by sentence.

2. Materials

To reveal the characteristics of between-sentence pauses in news speech, the speech uttered by newscasters was used as materials. The speech was newly recorded by the current authors.

10.21437/SpeechProsody.2020-139

2.1. Speakers

The speakers are professional newscasters (A: 30s, female; B: 20s, female) belonging to "the Nippon Television Network Corporation," which is one of the major TV stations in Japan. They have sufficient experience as newscasters. Both newscasters read the same manuscripts with some practice time in advance.

2.2. Reading manuscripts

We used 23 reading manuscripts, which were judged by a TV station as being representative for news style. Each manuscript was prepared on the assumption that it would be read out in about one minute. These reading manuscripts have been used in actual TV broadcasting. The covered topics were such as the expansion plan of the Narita Airport and a disaster prevention training utilizing augmented reality technology.

An interview was conducted with the newscasters to understand the actual situation of reading news manuscripts, and the followings were revealed.

- Reading news manuscripts is different from daily conversation as well as speaking situations such as program advertisements and narrations.
- When reading a manuscript, they always check not only the concerned part but also a following part, which is sometimes quite far away.
- They usually practice reading news manuscripts except for urgent news.
- In this reading practice, they try to find an appropriate way to control pauses and intonation, for easier understanding by the audience.
- Specifically, they write notes on their manuscript sheets such as the location and the degree of the length of pauses, connecting and separating words.
- Manuscripts containing these notes are also used in actual TV broadcasting.

Analysis and modeling of pauses described in the later chapters will take these facts into account.

2.3. Between-sentence pauses

A between-sentence pause is defined as a pause between sentences. The mean, standard deviation, minimum, and maximum of the number of occurrences of between-sentence pauses in a news manuscript were 4.4, 1.0, 3.0, 6.0, respectively.

Considering that pauses are influenced by physiological and linguistical factors in an elaborate way within an individual [1-4, 6-7], analysis and modeling of pauses are conducted for each newscaster in this study.

3. Extraction of acoustic features

The duration, F0, and intensity in phoneme units were extracted as acoustic features for analysis. The duration of each phoneme and pause was automatically measured by using the HTK toolkit. In this toolkit, speech is segmented with 10 msec resolution.

For the pitch-related parameters, the F0 values were computed every 10 msec by a conventional autocorrelation-based method [13]. All the F0 values were then converted to a musical (log) scale before any subsequent processing. The following equation was used to produce F0 values in semitone intervals.

$$F0[semitone] = 12*\log_2(F0[Hz])$$

The intensity of the speech signal was also calculated in dB for every 10 msec.

4. Analysis of the characteristics of between-sentence pauses

To clarify the characteristics of between-sentence pauses, their durations were analyzed. Moreover, relationship of between-sentence pause durations with prosodic features of the preceding and subsequent speech were analyzed.

4.1. Between-sentence pause duration

Between-sentence pause durations were measured. The results are shown in Figure 1.

The mean and standard deviation of duration were 1390.5 and 354.5 msec for newscaster A, and 1729.0 and 520.5 msec for newscaster B, respectively.

Then, the correlation between between-sentence pause durations by newscaster A and B for all reading manuscripts was investigated. Correlation coefficient was 0.49. The correlation coefficient increased to 0.57 by excluding four samples of the larger outliers (4240, 3490, 3340, and 2890 msec) of newscaster B. These samples were the between-sentence pauses after the first sentence of each reading manuscript. These sentences were lead sentences. Lead sentences are sentences which show abstract of each news and are located at the beginning of each manuscript. We may need a special treatment for them. Actually, these pauses are much longer than others, thus they make the mean of newscaster B longer than that of newscaster A.

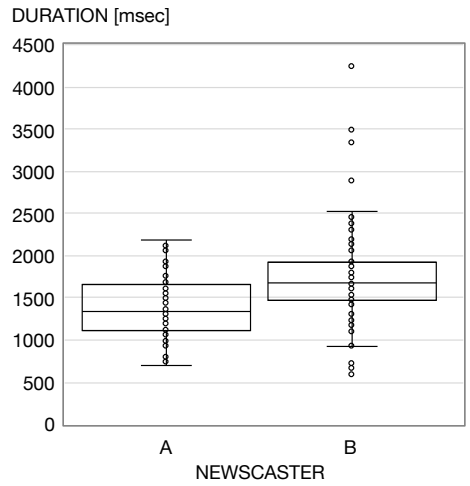

Figure 1: *The between-sentence pause duration in a news manuscript for two newscasters.*

Table 1: *Correlation of between-sentence pause durations with normalized prosodic features of the preceding and subsequent speech.*

PROSODIC FEATURE	CORRELATION COEFFICIENT Newscaster A		Newscaster B	
BEFORE the concerned between-sentence pause				
DURATION-related				
Duration of the preceding sentence	-0.12		0.10	
The number of morae included in the preceding sentence	0.11		-0.10	
F0-related				
Mean F0 of the last 3 morae in the preceding sentence	-0.35	***	-0.42	***
Mean F0 of the last 6 morae in the preceding sentence	-0.37	***	-0.42	***
Mean F0 of the last 10 morae in the preceding sentence	-0.18		-0.25	***
INTENSITY-related				
Mean intensity of the last 3 morae in the preceding sentence	-0.07		-0.10	
Mean intensity of the last 6 morae in the preceding sentence	-0.16		-0.14	
Mean intensity of the last 10 morae in the preceding sentence	-0.08		-0.10	
AFTER the concerned between-sentence pause				
DURATION-related				
Duration of the subsequent sentence	-0.11		0.17	*
The number of morae included in the subsequent sentence	0.24	***	0.20	***
F0-related				
Mean F0 of the first 3 morae in the subsequent sentence	0.07		0.10	
Mean F0 of the first 6 morae in the subsequent sentence	0.08		0.09	
Mean F0 of the first 10 morae in the subsequent sentence	0.03		0.07	
INTENSITY-related				
Mean intensity of the first 3 morae in the subsequent sentence	-0.05		0.05	
Mean intensity of the first 6 morae in the subsequent sentence	0.07		0.05	
Mean intensity of the first 10 morae in the subsequent sentence	0.08		0.07	

Significant difference ***: $p<0.001$ **: $p<0.01$ *: $p<0.05$

Prosodic features used for the model I: underlined

II: *italicized*

4.2. Correlation of between-sentence pause durations and prosodic features

Correlation of between-sentence pause durations and prosodic features of the preceding and subsequent speech was analyzed. Considering the information obtained in the interview with the newscasters described in Section 2.2, the prosodic features listed in Table 1 were utilized for analysis. These features are considered to reflect features of sentences/parts before/after the concerned between-sentence pause.

Table 1 also shows the results of correlation analysis of between-sentence pause durations and normalized prosodic features. Main features correlating with between-sentence pause durations were as follows.

a) F0-related features before the concerned between-sentence pause
b) Duration-related features after the concerned between-sentence pause

F0 of the last part in the preceding sentence was found to be negatively correlated with the between-sentence pause duration. It suggests that a lower F0 is an indicator of its following larger boundary, i.e., its following longer between-sentence pause duration. The number of morae included in the subsequent sentence was found to be positively correlated with the between-sentence pause duration. It suggests that a longer between-sentence pause duration is required to prepare for its following longer sentence. These pause durations may also contribute to understanding by listeners.

5. Modeling of between-sentence pause durations

To predict between-sentence pause durations, its modeling was attempted. Linear multiple regression model was used for prediction, with the between-sentence pause duration as a dependent variable and prosodic features mentioned in Section 4.2 as independent variables. The z-normalized values for each manuscript were utilized for the variables.

For the model I, prosodic features with significant differences underlined in Table 1 were selected as independent variables. Since "the mean F0 of the last 3 morae in the preceding sentence" and "the mean F0 of the last 6 morae in the preceding sentence" are similar, the latter showing higher correlation for both newscasters A and B was adopted. For a duration-related prosodic feature after the concerned between-sentence pause, the number of morae included in the subsequent sentence showing higher correlation for both newscasters A and B was adopted. The correlation coefficient between predicted and measured values by 3-fold cross-validation method was 0.49 for training data and 0.44 for test data, as shown in Table 2. Although the accuracy of this prediction model is not

Table 2: *The accuracy of prediction models.*

MODEL	VARIABLE	CORRELATION COEFFICIENT between predicted and measured between-sentence pause durations			
		TRAINING		TEST	
		Newscaster A	Newscaster B	Newscaster A	Newscaster B
I	Mean F0 of the last 6 morae in the preceding sentence The number of morae included in the subsequent sentence	0.51	0.54	0.45	0.43
II	*Duration of the preceding sentence* *The number of morae included in the preceding sentence* *Mean F0 of the last 6 morae in the preceding sentence* *Mean intensity of the last 6 morae in the preceding sentence* *Duration of the subsequent sentence* *The number of morae included in the subsequent sentence* *Mean F0 of the first 6 morae in the subsequent sentence* *Mean intensity of the first 10 morae in the subsequent sentence*	0.39	0.38	0.24	0.20

practical, it was found that between-sentence pause durations could be predicted to some extent by utilizing prosodic information of the preceding and subsequent speech.

For the model II, all prosodic features listed in Table 1 were used as independent variables, where each variable showing the strongest correlation was selected from "F0-related/intensity-related" features of "before/after the concerned between-sentence pause." The prosodic features used for the model II are italicized in Table 1. The correlation coefficient between predicted and measured values by 3-fold cross-validation method was 0.37 for training data and 0.23 for test data, as shown in Table 2. The accuracy of this prediction model with more variables was lower than that of model I. This result indicates that it is important to select effective variables, taking into consideration the relationship with the between-sentence pause duration.

6. Conclusions

Aiming at improving the news speech likeness of synthesized speech, we analyzed the characteristics of between-sentence pause durations of news speech by newscasters and constructed a model to predict these durations.

First, to clarify the characteristics of between-sentence pauses, the durations were analyzed. The mean and standard deviation of duration were 1390.5 and 354.5 msec for newscaster A, and 1729.0 and 520.5 msec for newscaster B, respectively.

Then, the correlation coefficient of between-sentence pause durations by newscaster A with those by B for all reading manuscripts was 0.49. The correlation coefficient increased to 0.57 by excluding four samples of the larger outliers of newscaster B. These samples had the common feature that they were taken from right after lead sentences. Namely, in such an environment, it was found that there is a high degree of freedom in adjusting between-sentence pause durations of news speech.

Then, relationships of the between-sentence pause durations with the prosodic features of the preceding and subsequent speech were analyzed. As a result, it became clear that the following prosodic features have higher correlation with the between-sentence pause durations: the mean F0 of the last six morae in the preceding sentence, and the number of morae included in the subsequent sentence.

The correlation coefficient between the predicted values by the linear multiple regression model using these parameters and the measured values was 0.44 for the test data. This accuracy was higher than that by the model using all prosodic features. In other words, it was shown that higher accuracy was obtained with fewer variables.

To analyze more detailed characteristics of the between-sentence pause duration and improve the accuracy of the prediction model of between-sentence pause durations, we will consider the structure peculiar to news manuscripts mentioned in Section 4.1, and semantic factors of each manuscript in the future.

7. Acknowledgements

This study was performed in collaboration with the Nippon Television Network Corporation. This study was supported by JST ERATO Ishiguro Symbiotic Human-Robot Interaction program (#JPMJER1401) and JSPS Grant-in-Aid for Scientific Research on Innovative Areas (#8105).

8. References

[1] B. Zellner, Pauses and the temporal structure of speech – Fundamentals of Speech Synthesis and Speech Recognition. E. Keller, Ed. Chichester: John Wiley & Sons, pp. 41–62, 1994.

[2] A. Cutler, K. Demuth, and J. McQueen, "Universality versus language specificity in listening to running speech," J. Psychological Science, vol. 13, no. 3, pp. 258–262, 2002.

[3] M. Denny, and A. Smith, "Respiratory control in stuttering speakers: evidence from respiratory high-frequency oscillations," J. Speech, Language, and Hearing Research, vol. 43, no. 4, pp. 1024-1037, 2000.

[4] J. P. Lund, and A. Koltab, "Brainstem circuits that control mastication: Do they have anything to say during speech?," J. Communication Disorders, vol. 39, no. 5, pp. 381–390, 2006.

[5] C. Männel et al., "The role of pause as a prosodic boundary marker: Language ERP studies in German 3- and 6-year-olds," J. Developmental Cognitive Neuroscience, vol. 5, pp. 86–94, 2013.

[6] S. Hiki, “Durational characteristics of various segments in continuous speech,” J. Institute of Electronics, Information and Communication Engineers, vol. 50, no. 8, pp. 1485–1490, 1967 in Japanese.
[7] K. Kakita, and S. Hiki, “Durational characteristics of sentence-medial and sentence-final pauses in the production of a paragraph,” Proc. the 170th Meeting of the Acoustical Society of America, vol. 138, no. 3, p. 1944, 2015.
[8] F. Goldman–Eisler, Psycholinguistics: Experiments in spontaneous speech, New York: Academic Press, 1968.
[9] F. Goldman–Eisler, “Pauses, clauses, sentences,” J. Language and Speech, vol. 15, pp. 103–113, 1972.
[10] D. Duez, “Silent and non-silent pauses in three speech styles,” J. Language and Speech, vol. 25, no. 1, pp. 11–28, 1982.
[11] H. Fujisaki, S. Ohno, and S. Yamada, “Analysis of occurrence of pauses and their durations in Japanese text reading,” Proc. ICSLP, pp. 1387–1390, 1998.
[12] A. Imai, R. Ikezawa, N. Seiyama, A. Nakamura, T. Takagi, E. Miyasaka, and K. Nakabayashi, “An adaptive speech-rate conversion method for news programs without accumulating time delay,” IEICE Transaction on Fundamentals of Electronics, Communications and Computer Sciences, vol. J83-A. no. 8, pp. 935–945, 2000. in Japanese.
[13] C. T. Ishi, H. Ishiguro, and N. Hagita, “Automatic extraction of paralinguistic information using prosodic features related to F0, duration and voice quality,” J. Speech Communication, vol. 50, no. 6, pp. 531–543, 2008.