Proceedings of DiSS'05, Disfluency in Spontaneous Speech Workshop.
10–12 September 2005, Aix-en-Provence, France, pp. 169-172.

The effects of filled pauses on native and non-native listeners' speech processing

Michiko Watanabe[1], Yasuharu Den[2], Keikichi Hirose[3] & Nobuaki Minematsu[1]

[1]Graduate School of Frontier Sciences, University of Tokyo, Japan
[2]Faculty of Letters, Chiba University, Japan
[3]Graduate School of Information Science and Technology, University of Tokyo, Japan

Abstract

Everyday speech is abundant with disfluencies. However, little is known about their roles in speech communication. We examined the effects of filled pauses at phrase boundaries on native and non-native listeners in Japanese. Study of spontaneous speech corpus showed that filled pauses tended to precede relatively long and complex constituents. We tested the hypothesis that filled pauses biased listeners' expectation about the upcoming phrase toward a longer and complex one. In the experiment participants were presented with two shapes at one time, one simple and the other compound. Their task was to identify the one that they heard as soon as possible. The speech stimuli involved two factors: complexity and fluency. As the complexity factor, a half of the speech stimuli described compound shapes with long and complex phrases and the other half described simple shapes with short and simple phrases. As the fluency factor phrases describing a shape had a preceding filled pause, a preceding silent pause of the same length, or no preceding pause. The results of the experiments with both native and non-native listeners showed that response times to the complex phrases were significantly shorter after filled or silent pauses than when there was no pause. In contrast, there was no significant difference between the three conditions for the simple phrases, supporting the hypothesis.

1. Introduction

Spontaneous speech, unlike written sentences or speech read aloud from written text, is full of disfluencies such as filled pauses (fillers), repetitions, false starts and prolongations. It has been reported that about six per 100 words are disfluent in conversational speech in American English [7]. It has been found that every 13 words are disfluent among female speakers and every 10 words are disfluent among male speakers in Japanese presentations [8]. In spite of their abundance in everyday speech not many empirical studies have been conducted into their effects on speech communication either in native or non-native languages.

Three general views are possible about the effects of disfluencies on listeners.

1) Disfluencies disturb listeners.
2) Disfluencies neither harm nor help listeners.
3) Disfluencies are helpful to listeners.

In a native language, listeners hardly seem to be disturbed by disfluencies. Fox Tree [3] found in her experiments using identical word monitoring task that existence of repeated words in a sentence did not affect reaction times to target words immediately after repetitions. This suggested that repetition of words had no effect on speech processing of listeners. Using the same methodology Fox Tree [4] tested the effects of two types of fillers, "um" and "uh", on native listeners' comprehension in English and Dutch. The author used the term "fillers" to refer only to the voiced parts of filled pauses. In both languages the time that listeners needed to monitor target words were shorter when "uh" was present immediately before the target words than when it was digitally excised. However, no difference was found between the two conditions with "um". The results indicated that "uh" was helpful to listeners while "um" neither helped nor hindered comprehension. In any case, her experiment showed no negative effect of fillers on listeners.

In contrast with the effects of repetitions and fillers, listeners' reaction times to target words were longer when the target words were preceded by false starts than when the false starts were cut out, indicating that false starts had negative effects even on native listeners.

Regarding disfluencies in non-native languages, some researchers have argued that they are the main obstacle for listeners' perception and comprehension of speech. Voss [9] asked German subjects to transcribe a stretch of spontaneous English and analysed the transcripts. He found that nearly one third of all perception errors were connected with disfluencies. Misunderstanding was due to either misinterpreting disfluencies as parts of words or to misinterpreting parts of words as disfluencies. Fukao et al. [5] reported that international students studying at Japanese universities had difficulties in coping with disfluencies in lectures. They were sometimes not able to distinguish filled pauses from words and had problems in processing ungrammatical sentences, repairs, omissions or speech errors in lectures.

On the other hand, it has been claimed that disfluencies are sometimes helpful to listeners. Blau [1] compared non-native listeners' comprehension of monologues between three conditions: (1) normal speed, (2) modified to include extra three second pauses inserted, on average, every 23 words, and (3) with similar pauses filled with hesitations such as "well", "I mean", and "uh". Comprehension of the filled pause version was significantly better than that of the normal version and slightly better than the silent pause version. The results indicated that filled pauses sometimes helped comprehension of non-native listeners.

Summarizing the discrepant results of previous research about the effects of disfluencies on non-native listeners, Buck [2] argued that disfluencies, as well as silent pauses, which slowed down the speech rate, helped comprehension of non-native listeners as long as disfluencies were recognised as disfluencies. If listeners failed to recognize disfluencies as such, they could have detrimental effects. However, as studies with native listeners showed, word repetitions and fillers "um", which slowed down the speech rate, measured by the amount of linguistic information conveyed per unit time, neither helped nor hindered comprehension [3], [4]. Buck's argument

needs more empirical support and detailed analysis of various types of disfluencies at different locations.

In the present research we have examined the effects of filled pauses at phrase boundaries on native and non-native listeners' ability to process speech. It has been reported that filled pauses amount to about 70 % of the total disfluencies in Japanese [8]. The Japanese language seems to have a wider variety of filled pauses than English and Dutch. "Ano", "e", "eto" and "ma" have been listed as the most frequent fillers both in dialogues and monologues [6], [10].

Corpus based studies of spontaneous Japanese showed that filled pauses tended to appear more frequently before relatively long and complex constituents. Watanabe et al. [12] showed that the probability of filled pauses occurring at clause boundaries increased in a roughly linear manner when expressed as a function of the number of words in the following clauses. Watanabe [11] carried out a study on Japanese phrases sandwiched between silent pauses longer than 200 ms, "Inter Pausal Units (IPUs)", which tended to be units shorter than clauses. These IPUs contained significantly larger numbers of morae, words and phrases when they were immediately preceded by fillers than when they were not.

Based on these findings we have inferred that listeners are making use of this tendency in occurrence of filled pauses when they process speech. In the experiments described below we have tested the hypothesis that filled pauses bias listeners' expectations about the upcoming phrase toward a relatively long and complex one. We have tested this hypothesis with native speakers of Japanese in Experiment 1 and with non-native speakers of Japanese in Experiment 2.

2. Experiment 1

2.1. Outline

A pair of shapes in the same colour was presented side by side on a computer screen, one a simple shape (circle, triangle or square) and the other a compound shape (two arrows attached to a paired shape. See Fig. 1). One second after a visual stimulus had appeared speech referring to one of the two shapes was played. Participants were instructed to press a button corresponding to a shape being referred to as soon as possible. The instruction given to the participants was as follows (translated from Japanese): "A woman is asking to bring a paper decoration in a certain colour and a shape. Which one is she asking for? Two pictures of paper appear on the computer screen. Please press either a left or right mouse button corresponding to the paper that she is asking for as soon as possible."

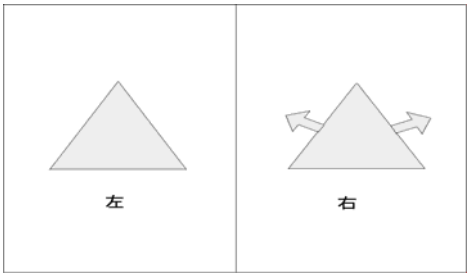

Figure 1: An example of visual stimuli. Visual stimuli always had a simple shape (round, square or triangular) on one side and a compound shape (with two arrows attached to the simple shape) on the other. The two shapes were always displayed in the same colour.

2.2. Speech stimuli

Each utterance contained a word describing a colour (we call it "a colour word") and a word describing a shape (we call it "a shape word") in this order as in "yellow and triangular". The speech stimuli involved two factors: 1) **complexity factor**: either a simple shape or a compound shape was referred to (we call the conditions, "simple condition" and "complex condition" respectively); 2) **fluency factor**: a shape word was immediately preceded by a filled pause, a silent pause of the same length as a filled pause, or no pause (we call the conditions, "filler condition", "pause condition" and "fluent condition", respectively). Examples of speech stimuli are given below with English translation. Fillers are in italic and phrases describing a shape are in bold.

An example of a simple phrase with a filler:
(1) Anone, tonari no heya kara kiirokute *eto* **sankaku no**
Look, next of room from yellow and *um* triangle of
kami mottekite kureru?
paper bring (auxiliary)
Translation: Look, could you bring a yellow and *um* **triangular** paper from the next room?

An example of a complex phrase with a filler:
(2) Anone, watashi no heya kara kiirokute *eto* **sankaku ni**
Look, I (genitive) room from yellow and *um* triangle to
yajirushi ga **tsuita** kami mottekite kureru?
arrows (nominative) attached paper bring (auxiliary)
Translation: Look, could you bring a yellow and *um* **triangular** paper **with arrows** from my room?

We assumed that filled pauses were more typical before a phrase describing a compound shape rather than before a phrase describing a simple shape because a phrase describing a compound shape was generally longer and more complex. We predicted that when a filler was uttered, listeners were more likely to expect a phrase describing a compound shape to follow. As a result, when a phrase describing a compound shape was actually uttered, listeners' response times to the phrase would be shorter than when there was no filler before the phrase. On the other hand, when a phrase describing a simple shape was uttered after a filler, listeners' response times to the phrase would not be shorter than when there was no filler because the filler was not in a typical location and therefore not a good cue to the type of phrase that followed.

We included the silent pause condition to examine whether silent and filled pauses of the same duration had different effects. As the other parts of the speech were kept constant, any difference should be attributable to whether the pause contained a voice or not.

Speech stimuli were created in the following way: one of the authors uttered sentences asking a supposed interlocutor to bring a sheet of paper of a certain colour and shape from a certain place. Although the test stimuli were presented to the speaker as a reading list, the speaker uttered sentences without looking at the list so that utterances sounded like natural, everyday speech. The speaker uttered 180 sentences. The utterances were recorded with an AKGC414B Studio microphone in an acoustically treated recording studio. The speech was sampled at 44 kHz and digitized at 16 bits directly onto a PC. All the utterances contained a filler "eto" immediately before a shape word. We called the original speech "a filler version". Original utterances were edited with speech analyzing software and two new versions were created: 1) a pause version: filled pauses were substituted by silence

with the same length as filled pauses; 2) a fluent version: filled pauses were edited out. Three sets of stimuli, each of which contained 180 sentences, were created so that only one of the three versions from the same utterances appeared in each stimuli set. The amplitude of speech stimuli was normalised.

2.3. Participants

Thirty university students who were native speakers of Tokyo Japanese took part in the experiment.

2.4. Procedure

The experiment was carried out in quiet rooms at Tokyo University and Chiba University in Japan. Participants were randomly assigned to one of the three stimuli sets. After eight practice trials the participants listened to 180 sentences. The order of stimuli was randomised for each participant. Speech stimuli were presented through stereo headphones. Sentences were played to the end no matter when the participants pressed a response button. Time out was set within 500 ms from the end of sentences. There were three second intervals between the trials. The experiment lasted about 40 minutes excluding the practice session and a short break in the middle.

Response times from the beginning of sound files were automatically measured. The onset of the first words describing a shape was marked manually referring to speech sound, sound waves and sound spectrograms. In the example sentences (1) and (2) the word onsets were marked at the beginning of /s/ in "**sankaku** (triangle)". Response times from the word onset were calculated by subtracting the word onset time from response times measured from the beginning of sound files.

The medians of correct response times from the word onset in each condition for each participant were calculated and the mean medians of six conditions were compared.

2.5. Results

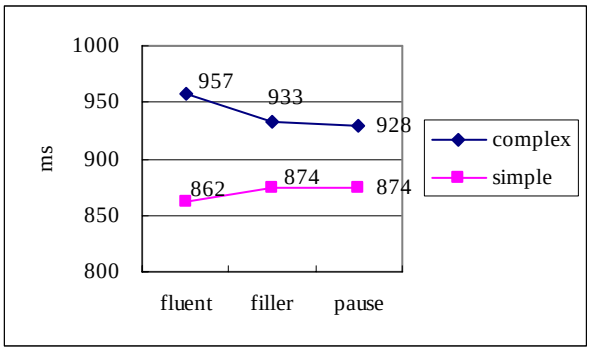

Figure 2: Japanese participants' mean response times from the onset of the first word describing a shape. 'Complex' means complex phrases and 'simple' means simple phrases.

Mean response times from the onset of shape words are shown in Figure 2. Two-way repeated measures analysis of variance (ANOVA) showed a main effect of complexity factor ($F(1, 29) = 76.051$, $p < .001$). A complexity-fluency interaction was significant ($F(2, 58) = 5.537$, $p < .006$). Post-hoc tests revealed that there was a significant difference between fluent-filler-pause conditions in the complex condition ($F(2, 28) = 6.533$, $p < .005$), but no significant difference in the simple condition ($F(2, 28) = 1.208$, $p = .314$). In the complex condition paired comparisons (alpha adjusted Bonferroni) showed significant differences between fluent-filler and fluent-pause conditions but no significant difference between filler-pause conditions ($t(29) = 3.329$, $p < .007$; $t(29) = 3.031$, $p < .015$; $t(29) = 0.492$, $p = 1.000$, respectively).

2.6. Discussion

Response times to complex phrases were shorter when a filled pause was present immediately before the phrase than when there was no preceding pause. On the other hand, there was no significant difference in response times to simple phrases between the filler and the fluent conditions. These results showed that existence of filled pauses accelerated listeners' responses to complex phrases but did not affect their responses to simple phrases, which was in accordance with our prediction and supported the hypothesis.

There was no significant difference in response times between the filler and the pause conditions in either the simple or the complex condition. This result indicated that the effects of filled pauses at phrase boundaries did not differ from the effects of silent pauses as long as the durations were the same.

3. Experiment 2

3.1. Outline and material

The outline and the material were the same as Experiment 1.

3.2. Participants

Thirty-eight native speakers of Chinese who had been staying in Japan for more than half a year and were fluent in everyday Japanese took part in the experiment. All the participants were either students or researchers at Chiba University or Tokyo University in Japan. Data from three participants were excluded from the analysis because they turned out to be bilingual speakers of Chinese and other languages. If the number of error trials, combined with trials which timed out, exceeded 18 for any participant, (i.e. exceeded 10% of presented trials), participants were excluded. That is, only participants scoring at least 162 out of 180 trials correct were considered for analysis. Five participants were excluded for this reason. This means that 30 participants were retained for analysis.

3.3. Procedure

The procedure was basically the same as that of Experiment 1. Most participants received eight practice trials, which was the same number presented to participants in Experiment 1. However, some participants did not press the response button until the utterance came to the end in the trial session. In each of these cases the participants were instructed not to wait until the speech ended, but to press the response button as soon as they knew the answer. They did four additional practice trials before starting the experiment.

3.4. Results

The ratio of the sum of errors and time outs of valid data was 2.6%. Mean response times of correct answers from the onset of shape words are shown in Figure 3.

Two way repeated measures analysis of variance (ANOVA) revealed that there were significant main effects of the complexity factor and the fluency factor ($F(1, 29) = 8.555$, $p < .007$; $F(2, 58) = 5.155$, $p < .009$, respectively). There was a significant complexity-fluency interaction ($F(2, 58) = 6.274$, $p < .003$). Post-hoc tests revealed that there was a significant difference between fluent-filler-pause conditions in the complex condition ($F(2, 28) = 7.867$, $p < .002$), but no significant difference in the simple condition ($F(2, 28) = .957$, $p = .396$). In the complex condition paired comparisons (alpha adjusted Bonferroni) showed significant differences between fluent-filler and fluent-pause conditions but no significant

difference between filler-pause conditions ($t(29) = 3.793$, $p < .002$; $t(29) = 3.219$, $p < .009$; $t(29) = 1.631$, $p = .341$, respectively).

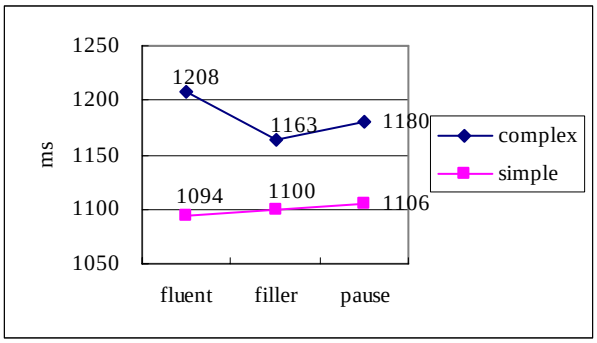

Figure 3: Chinese participants' mean response times from the onset of the first word describing a shape. 'Complex' means complex phrases and 'simple' means simple phrases.

3.5. Discussion

Although it took Chinese speakers 237ms longer on average to respond to correct figures than Japanese speakers, response times of Chinese speakers in the six conditions showed a similar pattern to those of Japanese speakers. Response times to complex phrases were shorter when a filled or silent pause preceded the phrase than when there was no preceding pause, while there was no significant difference in response times to simple phrases between any conditions. The results indicated that listeners tended to expect a longer and complex phrase to follow when they heard a filled pause, supporting the hypothesis.

The results agreed with Blau [1]'s results in that fillers, as well as silent pauses, helped non-native listeners' speech processing. Our results indicated that advanced language learners had acquired native like strategies in processing filled pauses.

4. Conclusion

The present research showed that filled pauses before long and complex phrases helped both native and non-native listeners' processing of speech by indicating complexity of the following phrase. Although our research was limited to the effects of one type of filler "eto", the results demonstrated that filled pauses at phrase boundaries were not harmful, at worst, and sometimes helpful to listeners. This is in accordance with the results of Fox Tree [4]'s study on fillers in English and Dutch. Our research provided information about the contexts in which filled pauses were helpful to listeners.

Our study with Chinese subjects suggested that advanced language learners were processing filled pauses in a way similar to native speakers. Namely, filled pauses at phrase boundaries seemed useful for non-native listeners as well as native listeners to predict complexity of the following phrase.

5. Acknowledgment

We thank Prof. Max Coltheart and Dr. Sallyanne Palethorp at Macquarie University for their kind advice in planning the experiment. This study was partly supported by JST/CREST the Expressive Speech Processing project.

6. References

[1] Blau, Eileen, Kay. 1991. More on comprehensible input: The effect of pauses and hesitation markers on listening comprehension, from ERIC database. Paper presented at the Annual Meeting of the Puerto Rico Teachers of English to Speakers of Other Languages (San Juan, PR, November 15, 1991).

[2] Buck, Gary, 2001, *Assessing Listening*, Cambridge: Cambridge University Press.

[3] Fox Tree, Jean Eleonore, 1995, The effects of false starts and repetitions on the processing of subsequent words in spontaneous speech. *Journal of memory and language* 34, pp. 709-738.

[4] Fox Tree, Jean Eleonore, 2001. Listeners' uses of *um* and *uh* in speech comprehension. *Memory and Cognition*, 29 (2), pp. 320-326.

[5] Fukao, Yuriko, Sumiko Mizuta & Kazuo Ohtsubo. 1991. Development of teaching material for advanced learners of Japanese to improve their listening skills for lecture comprehension. Paper presented at the autumn meeting of the Society for Teaching Japanese as a Foreign Language (in Japanese).

[6] Murakami, Jinich. & Shigeki Sagayama. 1991. A discussion of acoustic and linguistic problems in spontaneous speech recognition, *Technical report of IEICE*, SP91-100, NLC91-57: pp.71-78 (in Japanese).

[7] Shriberg, Elizabeth. 1994. *Preliminaries to a theory of speech disfluencies*. Ph.D. thesis, University of Berkeley, California.

[8] The National Institute for Japanese Language. *The Corpus of Spontaneous Japanese* homepage. http://www2.kokken.go.jp/%7Ecsj/public/6_1.html

[9] Voss, B. 1979. Hesitation phenomena as sources of perceptual errors for non-native speakers. *Language and Speech*, 22(2): pp. 129-44.

[10] Watanabe, Michiko. 2001. An analysis of usage of fillers in Japanese Lecture-style speech, *Proc. the Spontaneous Speech Science and Technology Workshop*, Tokyo. pp. 69-76 (in Japanese).

[11] Watanabe, Michiko. 2003. The constituent complexity and types of fillers in Japanese. *The Proc. of the 15th ICPhS*, pp. 2473-2476, Barcelona, Spain.

[12] Watanabe, Michiko, Yasuharu Den, Keikichi Hirose & Nobuaki Minematsu. 2004. Types of clause boundaries and the frequencies of filled pauses. *Proc. the 18th General Meeting of the Phonetic Society of Japan.* pp. 65-70. (in Japanese).