Filled Pauses as Cues to the Complexity of Following Phrases

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

[1] Graduate School of Frontier Sciences, University of Tokyo, Japan
[2]Graduate School of Information Science and Technology, University of Tokyo, Japan
[3]Faculty of Letters, Chiba University, Japan
watanabe@gavo.t.u-tokyo.ac.jp

Abstract

Corpus based studies of spontaneous speech showed that filled pauses tended to precede relatively long and complex constituents. We examined whether listeners made use of such a tendency in speech processing. We tested the hypothesis that when listeners heard filled pauses they tended to expect a relatively long and complex phrase to follow. In the experiment participants listened to sentences referring to both simple and compound shapes presented on a computer screen. Their task was to press a button as soon as they had identified the shape that they heard. The sentences involved two factors: complexity and fluency. As the complexity factor, a half of the sentences 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 as the filled pause, or no preceding pause. The results showed that response times for 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. The results support the hypothesis and indicate that it is the duration of filled pauses that give listeners cues to the complexity of upcoming phrases.

1. Introduction

Everyday speech is full of disfluencies such as filled pauses (fillers), repetitions, false starts, and prolongations. It has been reported that the disfluency ratio is about six per 100 words in conversational speech in English [1]. In spite of their abundance in everyday speech only a small number of empirical studies has been conducted into their roles in speech communication.

As disfluencies hardly appear in speech that is read aloud from written text, they are considered to be relevant to on-line speech planning: when people have some difficulty in the timing constraints underpinning their speech planning and execution, they may be disfluent. Major constituents such as sentences, clauses, noun phrases (NP), verb phrases (VP), and prepositional phrases (PP) have been regarded as principal units of planning [2], [3], [4], [5]. It has been found that disfluencies tend to appear more frequently at the beginning of these constituents rather than in the other positions [1], [6].

It has been reported that the longer and more complex the following constituents the more frequent disfluencies. Clark and Wasow [6] found that repetition rates of articles at the beginning of NPs were higher when the NPs were longer and more complex as well as when the NPs were located closer to the beginning of a sentence and therefore the numbers of constituents following the NPs were larger. Watanabe et al. [7] investigated the ratios of filled pauses after four types of case particles in Japanese. Case particles are located at the end of NPs in Japanese. The authors found that the closer the case tended to be located to the beginning of a sentence, the higher the ratio of filled pauses: the ratio of filled pauses was highest after topic particles, next highest after nominative particles, and lowest after dative and accusative particles. The order of the ratios of filled pauses after four types of case particles corresponded to the order of length and complexity of the constituents following the case.

It has been pointed out that constituents tend to be longer and more complex when the constituents are preceded by filled pauses than when they are not. In talks about given topics by 10 subjects, Cook, Smith and Lalljee [8] found that the number of words in clauses immediately following filled pauses was significantly larger than the number of words in clauses which were not preceded by filled pauses. However, when the same subjects described and summarized cartoons without captions, the number of words in clauses following filled pauses was not significantly larger than the number of words in clauses which were not preceded by filled pauses. The authors speculated that the inconsistent results of the two types of speech were attributable either to their differences in syntactic complexity or in the number of samples due to different durations. Watanabe et al. [7] reported that a study of pauses in academic and casual presentations in Japanese showed that clauses immediately after filled pauses contained more words than clauses not preceded by filled pauses, which confirmed the tendency found by [8]. Watanabe [9] carried out a study on Japanese phrases sandwiched between silent pauses longer than 200 ms, an Inter Pausal Unit (IPU), which tended to be units shorter than clauses. These IPU contained significantly larger numbers of morae, words and phrases when they were immediately preceded by fillers than when they were not; this added evidence that constituents tended to be longer and more complex when the constituents were preceded by filled pauses than when they were not.

These findings support the assumption mentioned earlier that disfluencies are related to the degree of speakers' difficulty in planning the following constituents. Clark and Wasow [6] argued that the degree of difficulty in planning the following constituents corresponded to the "grammatical weight" of the constituents. "Grammatical weight" was defined as "roughly the amount of information expressed in a constituent" and was referred to as "complexity" [6]. Grammatical weight, or complexity, was measured by the number of words, syntactic nodes, or phrasal nodes in the constituent and these numbers were reported to correlate with

each other at .94 and beyond [10]. Therefore, all were supposed to be equally useful as measures of complexity.

In the present research we have examined whether listeners are making use of this tendency in distribution of filled pauses in speech processing. Before describing our experiment, we briefly survey research into effects of disfluencies on listeners. Three general views are possible about effects of disfluencies on listeners.

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

Previous research has indicated that effects of disfluencies depend on the type and the location of them. In Fox-Tree's experiments using identical word monitoring task [11] listeners' reaction times for target words were longer when the target words were preceded by false starts than when the false starts were digitally excised, indicating that false starts have negative effects on listeners. And the negative effects were more obvious when false starts occurred in the middle of sentences than when they did at the beginning of sentences, showing that the location of false starts also affects the effects. In contrast with false starts, existence of repetitions did not affect reaction times to target words immediately after repetitions, suggesting that repetitions have neither detrimental nor beneficial effects on comprehension.

Using the same methodology, Fox Tree [12] tested effects of two types of fillers, "um" and "uh", on comprehension in English and Dutch. Fox Tree used the term "fillers" to refer only to the voiced parts of filled pauses. In the both languages the time that listeners needed to monitor target words were shorter when "uh" was present immediately before the words than when it was digitally excised, but no difference was found between the conditions when "um" was present and when it was not. The results indicate that "uh" helps listener comprehension while "um" neither helps nor hinders comprehension. The author argues that "uh" helps comprehension by signalling a short delay and heightening listeners' attention for upcoming speech whereas "um", which signals a longer delay, does not have such a function.

In the present research we have tested the hypothesis that listeners tend to expect a relatively long and complex phrase to follow when they hear a filled pause. We composed utterances in each of which either a simple or a complex phrase appeared. Some of the phrases were preceded by a filled pause, some were preceded by a silent pause, and the others had no preceding pause. We have predicted that listeners' response to a complex phrase will be quicker when the phrase is preceded by a filled pause than when the phrase has no preceding pause, because a filled pause signals the speaker's difficulty in speech planning and allows listeners time to predict the content of the upcoming speech.

We have added the silent pause condition to examine whether filled pauses have different effects from silent pauses. It has been known that silent pauses at constituent boundaries help listener comprehension by providing listeners with time for processing speech [13], [14], [15]. On the other hand it has been reported that silent pauses do not have apparent beneficial effects on listeners if the speech itself or the tasks that listeners have to do after hearing the speech are easy enough [16]. Like silent pauses, filled pauses provide listeners with extra time. On the other hand, unlike silent pauses, filled pauses contain a sound. If there is a different effect on listeners of filled pauses of the same length as silent pauses, the effect should be attributable to the sound in filled pauses.

It is predicted from the hypothesis that filled pauses before a simple phrase can have a detrimental effect. As filled pauses more frequently appear before a long and complex phrase, listeners may expect a wrong type of phrase to follow. Consequently, listeners' response to a simple phrase can be slower when there is a preceding filled pause than when there is no pause. However, as filled pauses do sometimes occur before simple phrases, they may not have any negative effects. As filled pauses allow listeners extra time like silent pauses, they may benefit listeners to certain extent even at atypical positions.

2. Experiment

2.1. Outline

A pair of coloured shapes 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). We assumed that it would take a speaker longer to plan a phrase describing a compound shape than to plan a phrase describing a simple shape because the speaker would need more words and more complex syntactic structures for a compound shape than for a simple shape. 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."

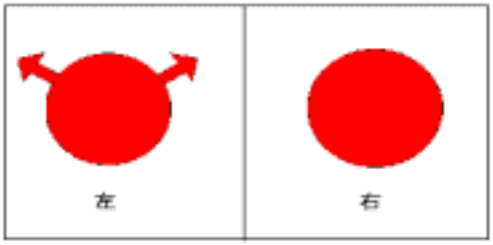

Figure 1: An example of visual stimuli. Visual stimuli always had a simple shape (round, square or triangular) on one side and a compound shape (with two arrows attached to the simple shape) on the other. The two shapes were always displayed in the same colour.

The speech always contained a word describing a colour (a colour word) and a word describing a shape (a shape word) in this order. One third of the speech stimuli contained a filled pause between a colour word and a shape word. One third of the speech stimuli contained a silent pause of the same length as a filled pause between a colour word and a shape word. One third of the speech stimuli contained neither a filled pause nor a silent pause between a colour word and a shape word. Our prediction was that response times to a complex phrase describing a compound shape would be shorter when

the phrase was preceded by a filled pause than when there was no pause. Response times from the onset of the first word describing a shape were measured.

2.2. Speech stimuli

The speech stimuli involved two factors: 1) describing a simple shape or a compound shape (complexity factor); 2) a phrase describing a shape was preceded by a filled pause, a silent pause of the same length as a filled pause, or no pause (fluency factor). Examples of speech stimuli are given below with English translation. Fillers are in italic and phrases describing a shape are in bold. When a filled pause was in an utterance, it always appeared between a colour word and a shape word.

An example of a complex phrase with a filler:
(1) Anone, watashi no heya kara akakute *eto* **maru ni**
Look, I (genitive) room from red and *um* circle to
yajirushi ga **tsuita** kami mottekite kureru?
arrows (nominative) attached paper bring (auxiliary)
Translation: Look, could you bring a red and *um* **round** paper **with arrows** from my room?

An example of a simple phrase with a filler:
(2) Anone, tonari no heya kara akakute *eto* **marui** kami
Look, next (genitive) room from red and *um* circular paper
mottekite kureru?
bring (auxiliary)
Translation: Look, could you bring a red and *um* **round** paper from the next room?

Speech stimuli were created in the following way: one of the authors uttered sentences asking a supposed interlocutor to bring a sheet of paper in a certain colour and a shape at 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 using 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. Procedure

Thirty university students who were native speakers of Tokyo Japanese took part in the experiment. 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 /m/ in "**maru** (circle)" and "**marui** (circular)" respectively. 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.

2.4. Results

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).

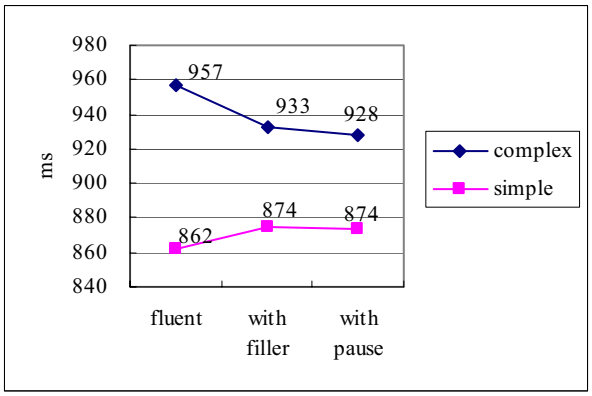

Figure 2: Mean response times from the onset of the first word describing a shape. 'Complex' means complex phrases and 'simple' means simple phrases.

3. Discussion

In the complex condition response times for the filler condition were shorter than those for the fluent condition. For simple phrases there was no significant difference in response times between any conditions. Filled pauses before simple phrases showed neither positive nor negative effects. This would suggest that filled pauses facilitated the speech processing of listeners only when they preceded complex

phrases. These results support the hypothesis that listeners tend to expect a relatively long and complex phrase to follow when they hear filled pauses.

In the complex condition response times for the silent pause condition as well as the filled pause condition were significantly shorter than those of the fluent condition, whereas no significant difference was found between the two conditions in the simple condition, suggesting that silent pauses had positive effects only before a complex phrase. There was no significant difference in response times between silent pause and filled pause conditions either in the simple or the complex condition. Therefore, silent pauses of the same length as filled pauses had the same effects on response times as filled pauses both in the simple and the complex conditions. This result indicates that it is the time that filled pauses take that helped listeners respond to a complex phrase, rather than the sound contained in them.

It has been claimed that silent pauses at constituent boundaries help listener comprehension by giving listeners time to process the speech up to that point. If it had been the case with our experiment, response times after a silent pause should have been shorter than response times without a pause in the simple condition as well as in the complex condition because phrases preceding a silent pause were exactly the same in the simple and the complex conditions. Therefore, different effects of silent pauses in the simple and the complex conditions suggest that listeners used the pauses to predict the following parts rather than to process the previous parts. In our experiment the word immediately before a target word was always a colour word. As two shapes on a computer screen always had the same colour, it was not necessary for listeners to process a colour word to do the task. This also indicates that listeners tended to use pauses for predicting the upcoming parts rather than for processing the previous parts.

Filled pauses before a simple phrase could have had a negative effect on response times because they were not in a typical location, but they did not. Response times to a simple phrase were constantly shorter than those to a complex phrase, which suggests that phrases describing a simple shape were easier for listeners to process than phrases describing a compound shape. Filled pauses before a simple phrase did not have a negative effect, possibly because phrases describing a simple shape were easy enough for listeners to process without delay even with a potentially detrimental factor. Another explanation would be that as filled pauses before a simple phrase were not typical but not uncommon either, they hardly biased listeners' expectation of the following phrase and had no obvious effects on response times. In any case, our results showed no negative effects of filled pauses on listeners' speech processing, which is in accordance with Fox-Tree [12]'s results on English and Dutch filled pauses. Our results indicate that filled pauses at constituent boundaries do not disturb listeners but sometimes help them predicting upcoming speech.

4. Conclusion

In the present research we have tested the hypothesis that listeners tend to expect a relatively long and complex phrase to follow when they hear a filled pause. Listeners' response times to a complex phrase were shorter when the phrase was preceded by a filled pause than when there was no pause before the phrase. On the other hand, no difference was found in response times to a simple phrase between the two conditions. These findings show that it is only the filled pause before a complex phrase that facilitates listeners' speech processing, supporting the hypothesis. Silent pauses had the same effects on listeners' response times as filled pauses of the same duration. This suggests that the time that silent and filled pauses take affects listeners' prediction about the following speech. Our results have provided evidence that filled pauses at phrase boundaries are not harmful, at worst, and sometimes helpful for listeners in processing upcoming speech.

5. References

[1] Shriberg, E., E., *Preliminaries to a Theory of Speech Disfluencies*, PhD thesis, University of California at Berkeley, 1994.

[2] Goldman-Eisler, F., *Psycholinguistics*, London: Academic Press, 1968.

[3] Holms, V. M., "Hesitations and sentence planning", *Language and cognitive processes*, 3(4): 323-361, 1988.

[4] Levelt, W. J. M., *Speaking,* The MIT Press: Cambridge, Massachusetts, 1989.

[5] Maclay, H. and Osgood, C. E., "Hesitation phenomena in spontaneous English speech", *Word,* 15, 19-44, 1959.

[6] Clark, H. H., & Wasow, T., "Repeating words in spontaneous speech", *Cognitive Psychology*, *37*: 201-242, 1998.

[7] Watanabe, M., Den, Y., Hirose, K., & Minematsu, N., "Clause types and filled pauses in Japanese spontaneous monologues", *Proc. of the 8th ICSLP*, 905-908, Jeju Island, Korea, 2004.

[8] Cook, M., Smith, J., & Lalljee, M. G., "Filled pauses and syntactic complexity", *Language and Speech*, 17:1, 11-16, 1974.

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

[10] Wasow, T., "Remarks on grammatical weight", *Language variation and change*, 9, 81-105, 1997.

[11] Fox Tree, J. E., "The effects of false starts and repetitions on the processing of subsequent words in spontaneous speech", *Journal of memory and language* 34, 709-738, 1995.

[12] Fox Tree, J. E., "Listeners' uses of *um* and *uh* in speech comprehension", *Memory and Cognition*, 29 (2), 320-326, 2001.

[13] Griffiths, R. "Pausological research in an L2 context: A rationale, and review of selected studies", *Applied Linguistics*, 12(4): 345-364, 1991.

[14] Reich, S. S., "Significance of pauses for speech perception", *Journal of Psycholinguistic Research,* 9(4), 379-389, 1980.

[15] Sugito, M., "On the role of pauses in production and perception of discourse", *Proc. of the 1st ICSLP* 1990, 513-516, Kobe, Japan, 1990.

[16] Aaronson, D., "Temporal course of perception in an immediate recall task", *Journal of Experimental Psychology*, 76, 129-140, 1968.