Available online at www.sciencedirect.com
ScienceDirect
Speech Communication 50 (2008) 81–94
SPEECH COMMUNICATION
www.elsevier.com/locate/specom

Filled pauses as cues to the complexity of upcoming phrases for native and non-native listeners

Michiko Watanabe [a,*], Keikichi Hirose [b], Yasuharu Den [c], Nobuaki Minematsu [a]

[a] Graduate School of Frontier Sciences, Engineering Building, No. 2, Room 103c2, The University of Tokyo, 7-3-1 Hongo, Bunkyo-ku, Tokyo 113-0033, Japan
[b] Graduate School of Information Science and Technology, The University of Tokyo, 7-3-1 Hongo, Bunkyo-ku, Tokyo 113-0033, Japan
[c] Faculty of Letters, Chiba University, 1-33 Yayoi-cho, Inage-ku, Chiba-shi, Chiba 263-8522, Japan

Received 21 February 2006; received in revised form 14 June 2007; accepted 14 June 2007

Abstract

We examined whether filled pauses (FPs) affect listeners' predictions about the complexity of upcoming phrases in Japanese. Studies of spontaneous speech corpora show that constituents tend to be longer or more complex when they are immediately preceded by FPs than when they are not. From this finding, we hypothesized that FPs cause listeners to expect that the speaker is going to refer to something that is likely to be expressed by a relatively long or complex constituent. In the experiments, participants listened to sentences describing both simple and compound shapes on a computer screen. Their task was to press a button as soon as they had identified the shape corresponding to the description. Phrases describing shapes were immediately preceded by a FP, a silent pause of the same duration, or no pause. We predicted that listeners' response times to compound shapes would be shorter when there is a FP before phrases describing the shape than when there is no FP, because FPs are good cues to complex phrases, whereas response times to simple shapes would not be shorter with a preceding FP than without. The results of native Japanese and proficient non-native Chinese listeners agreed with the prediction and provided evidence to support the hypothesis. Response times of the least proficient non-native listeners were not affected by the existence of FPs, suggesting that the effects of FPs on non-native listeners depend on their language proficiency.

Keywords: Filled pauses; Listeners; Prediction; Complexity; Japanese

1. Introduction

Everyday speech is abundant with disfluencies such as filled pauses (fillers), repetitions, non-lexical prolongations and false starts. About 6 per 100 words are disfluent in conversational speech in American English (Fox Tree, 1995; Shriberg, 1994). In Japanese, fillers alone consist of about 6% of the total words in presentations (The National Institute of Japanese Language, 2004). In spite of their abundance in everyday speech, a relatively small number of empirical studies have been conducted on their roles in speech communication, particularly in languages other than English and Dutch. In this study we investigate the effects of filled pauses, the most frequent type of disfluencies, on native Japanese and non-native (Chinese) listeners' speech processing. More specifically, we examine whether filled pauses affect listeners' expectations about the complexity of upcoming speech.

This article is organized in the following way. The present section contains a survey of research on the features and the roles of disfluencies for speakers and listeners with particular attention paid to filled pauses. We focus on the findings of previous research which assert that filled pauses and repetitions are frequent before relatively long or complex constituents. Based on these findings, we propose the hypothesis that filled pauses cause listeners to expect a relatively long or complex phrase. Section 2 describes an experiment to test the hypothesis with native speakers of Japanese. Section 3 reports the same experiment conducted

* Corresponding author. Tel.: +81 3 5841 6767; fax: +81 3 5841 6648.
E-mail addresses: watanabe@gavo.t.u-tokyo.ac.jp (M. Watanabe), hirose@gavo.t.u-tokyo.ac.jp (K. Hirose), den@cogsci.l.chiba-u.ac.jp (Y. Den), mine@gavo.t.u-tokyo.ac.jp (N. Minematsu).

0167-6393/$ - see front matter © 2007 Elsevier B.V. All rights reserved.
doi:10.1016/j.specom.2007.06.002

with non-native (Chinese) speakers of Japanese. Section 4 discusses the results, and Section 5 concludes the study.

1.1. Disfluencies and speech planning

As disfluencies hardly appear in speech that is read aloud from written texts, they are considered to be relevant to on-line speech planning: when people have some difficulty in the time constraints underpinning their speech planning and execution, they are likely to be disfluent (Clark, 2002).

Three main stages are assumed in speech planning, conceptualizing a message, formulating the appropriate linguistic forms, and articulating them (Levelt, 1989). These three stages can run in parallel for different speech units, and disfluencies may occur at any of these stages (Clark and Fox Tree, 2002).

Major constituents such as sentences, clauses, and phrases have been claimed to be principal units of planning (Goldman-Eisler, 1968; Holms, 1988; Levelt, 1989; Maclay and Osgood, 1959). Disfluencies, particularly filled pauses and repetitions, are more frequent at the beginning of these constituents rather than in the other positions (Clark and Wasow, 1998; Shriberg, 1994). This phenomenon is congruent with the assumption that disfluencies reflect speakers' difficulties in conceptual planning and linguistic encoding.

Previous research has shown that the longer or more complex a constituent, the higher the disfluency rate immediately before or at the beginning of the constituent. Shriberg (1994) reported that the more words a sentence contains, the higher the probability of the sentence being disfluent particularly at the beginning of it. Clark and Wasow (1998) found that the longer and more complex a noun phrase (NP), the higher the repetition rate of the article. Watanabe et al. (2004b) investigated the rate 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 to the beginning of a sentence the case particle is located, the higher the rate of filled pauses after the case particle: the rate of filled pauses was the highest after topic particles, the next highest was after nominative particles, and the lowest rate was after dative and accusative particles. The order of the rates of filled pauses after four types of case particles corresponded to the order of the length and the complexity of the constituents following the case particle.

It has also been reported that constituents tend to be longer or more complex when the constituents are preceded by filled pauses than when they are not. Cook et al. (1974) found that the mean number of words in clauses immediately after filled pauses was significantly larger than the number of words in clauses without preceding filled pauses in talks about given topics in English. Watanabe et al. (2004a) reported that clauses immediately after filled pauses contained more words than those without preceding filled pauses in presentations in Japanese. This finding is consistent with the results of Cook, Smith and Lalljee. Watanabe (2003) also compared lengths of phrases sandwiched between silent pauses longer than 200 ms, or "Inter Pausal Units" (IPUs) in Japanese presentations. IPUs contained significantly larger numbers of morae, words or phrases when they were preceded by filled pauses than when they were not. These findings support the assumption that disfluencies reflect speakers' difficulties and need of time in planning the following constituents.

1.2. Effects of disfluencies on native listeners

Disfluencies have been rather negatively viewed because they seem to interrupt communication and waste time for listeners. However, recent studies have argued that disfluencies can contribute to smooth communication by not only allowing speakers extra time for planning, but also informing listeners of the speakers' mental attitudes or planning difficulties. While listeners are waiting for the speaker to continue instead of taking a turn, they may infer the reason for the speaker's trouble, predict the upcoming utterance, and prepare for it or offer help to the speaker (Clark, 2002; Shriberg, 2005; Stenstroem, 1994). Filled pauses, for example, have been observed to be frequent in dispreferred responses or embarrassing remarks (Finegan, 1994; Rose, 1998; Sadanobu and Takubo, 1995). Such filled pauses are likely to hint at the direction or type of the following utterance and prepare listeners for rather unwelcome remarks.

Brennan and Williams (1995) investigated whether listeners are sensitive to speakers' meta-cognitive states expressed by silent and filled pauses. They found that listeners evaluated the speaker as less confident when the speaker answered general knowledge questions with filled pauses than without delay or with silent pauses of the same duration. The results suggest that filled pauses affect listeners' evaluations of the contents of the following utterance.

Other studies have observed that disfluencies more directly affect the language processing of listeners. Fox Tree (1995) examined the effects of false starts and word repetitions on listener comprehension of speech, using an identical word monitoring task. Reaction times to the target words were longer when they were preceded by false starts than when the false starts were digitally excised. On the other hand, the existence of repeated words before the target words did not affect reaction times. The results suggest that false starts disturb listener comprehension of speech, whereas word repetitions have neither positive nor negative effects.

Using the same methodology, Fox Tree (2001) examined the effects of two types of fillers, *um* and *uh* in English and Dutch on listener comprehension of speech. The author found positive effects with *uh*, but not with *um* in both languages. The results suggest that the effects of filled pauses may depend on their sound forms. However, no negative effects of fillers were observed in this study. Brennan and Schober (2001) found that filled pauses between discarded and repairing words such as *uh* in "yel- uh, purple" help lis-

teners compensate for disruptions and delays in speech. Listeners in their experiments responded to repairing words more quickly and accurately when there was a filled pause before the target words than when there was a shorter silent pause. The authors argued that such effects of filled pauses are due to the extra time which elapses during the filled pauses, not their sounds, because silent pauses of the same duration had the same effects as the filled pauses.

In contrast with the results of Brennan and Schober (2001), Fox Tree (2002) found that filled pauses at turn beginnings have different effects from those of silent pauses. The participants (overhearers) judged second speakers to have more serious production difficulties when their turns were preceded by a filler, *um*, than when there were silent pauses of the same duration. The results indicate that filled pauses at turn beginnings signal the next speakers' production difficulties more explicitly than silent pauses of the same duration.

Disfluencies are known to be more frequent before objects which are newly introduced in the discourse (''discourse-new'' objects) than before items which have already been introduced (''discourse-given'' objects) (Arnold et al., 2000). Based on this distributional tendency, Arnold et al. (2003) hypothesised that disfluencies bias listeners' expectations toward discourse-new objects. In the experiment, participants were asked to move objects on a computer screen following instructions such as ''Now put the candle below the salt shaker'', and their eye movements were tracked. Immediately after the onset of the determiner, the participants fixated their eyes on discourse-new objects longer than on discourse-given items when the instruction contained disfluencies such as prolonged *the* (pronounced as /ðiː/) followed by *um* or *uh* before the target word. The authors concluded that disfluencies bias listeners' expectations toward discourse-new items and thus affect core language processing.

Bailey and Ferreira (2003) found that filled pauses affect listeners' parsing of syntactically ambiguous sentences. When filled pauses were in the pre-nominal positions as in example 1 below, ''the deer'' tended to be assigned correctly as the subject of the following clause rather than the object of the preceding verb. On the other hand, when filled pauses were in the post-nominal positions as in example 2, ''the deer'' tended to be assigned as the object of the preceding verb rather than the subject of the following clause. Consequently, the rate at which listeners judged example 1 as grammatical was higher than the rate at which they judged example 2 as grammatical, although the two cases had the same semantic representations.

(1) While the man hunted the *uh uh* deer ran into the woods.
(2) While the man hunted the deer *uh uh* ran into the woods.

The authors discussed two possible mechanisms for these phenomena. One was as follows: Filled pauses delay the onset of the disambiguating word, ''ran'' in example 2, which allows the parser a long time to be committed to the ''wrong'' analysis. This causes the alternative analysis to lose activation and makes reanalysis difficult. As a consequence, example 2 is judged ungrammatical more frequently than example 1. In this view, the time that passes during the filled pauses is assumed to play a critical role. The other possibility mentioned was that listeners are making use of the information about syntactic contexts in which filled pauses are frequent. In example 1, most listeners correctly judged ''the'' before *uh* as the beginning of a new clause because the location of the filled pauses is consistant with their natural distribution. In the case of example 2, many listeners seem to have expected a clause boundary after ''deer'' because filled pauses are frequent immediately before a new clause. However, as no NP for the subject of a new clause followed filled pauses, the listeners who failed to reanalyse the sentence seem to have judged the sentence ungrammatical. The authors concluded that both mechanisms seem to be at work, and that filled pauses can affect the parsing of temporally ambiguous sentences.

1.3. Effects of disfluencies on non-native listeners

Although it has been claimed that disfluencies are directly relevant to research in foreign language learning and teaching, only a small number of empirical studies have been conducted about their effects on non-native listeners (Buck, 2001; Griffiths, 1991; Rose, 1998). Some researchers have argued that they are the main obstacles to listener perception and comprehension of speech. Voss (1979) asked German subjects to transcribe a stretch of spontaneous English and analysed the transcripts. He found that nearly one-third of the perception errors were connected with disfluencies. Misunderstanding was due to either misinterpreting disfluencies as parts of words or misinterpreting parts of words as disfluencies. Fukao et al. (1991) reported that international students studying at Japanese universities had difficulties in coping with disfluencies in lectures. Some students could not distinguish filled pauses from words, and others were not able to comprehend speech with repairs, omissions, or errors.

On the other hand, Blau (1991) claimed that disfluencies can help non-native listeners' comprehension of speech. Blau compared non-native listeners' comprehension of monologues in English under three conditions: (1) speech at normal speed, (2) speech with extra three second silent pauses, on average, between every 23 words, and (3) speech with similar pauses filled with hesitations such as ''well'', ''I mean'', and *uh*. In one experiment with Spanish speakers, comprehension scores of the silent pause and the hesitation pause versions were significantly higher than those of the normal version. There was no significant difference between the scores of the silent pause and the hesitation pause versions. In another experiment with Japanese speakers, comprehension scores of the hesitation pause version were

significantly higher than those of the silent pause and the normal versions. Blau's study suggests that discourse markers and filled pauses help non-native listener comprehension. However, exactly what made the hesitation pause version the most helpful to listeners is not clear, because neither the duration of the hesitation pauses nor the prosody of the speech surrounding them seem to have been controlled. Positive effects of hesitation pauses may have been due to their longer durations than silent pauses. Silent pauses of the same duration may have the same effects as hesitation pauses. Also, it may have been the prosody of the surrounding speech or the hesitation pauses combined with the prosody that caused the advantage rather than the hesitation pauses alone. Given that hesitation pauses themselves facilitated the listener comprehension of speech, whether both discourse markers and filled pauses are helpful to listeners is not known. Discourse markers may have a positive effect, while fillers may not, or vice versa. Even among discourse markers or filled pauses, some types may be helpful, while others may not.

Summarizing the discrepant results of previous research about the effects of disfluencies on non-native listeners, Buck (2001) argued that disfluencies that slow the speech rate down help the comprehension of non-native listeners as long as the disfluencies are recognized as disfluencies. If listeners fail to recognize the disfluencies as such, they can have detrimental effects. However, as studies with native listeners indicated, word repetitions, which slow down the speech rate measured by the amount of linguistic information conveyed per unit of time, neither helped nor hindered comprehension (Fox Tree, 1995). Buck's argument needs more empirical support and detailed analyses of various types of disfluencies at different locations.

1.4. Hypothesis and predictions

As mentioned in Section 1.1, disfluencies are more frequent before relatively long or complex constituents. This is likely to be because long or complex constituents are more difficult to formulate and take speakers longer to plan than short or simple constituents. In conversations, one alternates between one's roles as a speaker and a listener. Therefore, it is plausible to assume that what one experiences as a speaker affects one's behaviour as a listener. It is also possible that listeners are utilizing filled pauses as probabilistic cues to the onset of relatively large constituents, as Bailey and Ferreira (2003) argued. We hypothesized from the findings of corpus-based studies that listeners expect a relatively long or complex phrase to follow when there is a filled pause. In other words, we predicted that filled pauses cause listeners to expect that the speaker is going to refer to something that requires a relatively long or complex constituent to express. Although the length measured by the number of words in a constituent and its syntactic complexity are not logically connected, they are highly correlated (Wasow, 2002). Therefore, we did not distinguish these two factors in this study. We have chosen to examine the effects of filled pauses on listeners because filled pauses are far more common than any other type of disfluencies in Japanese (The National Institute of Japanese Language, 2004).

We composed utterances in which either a simple or a complex phrase appeared to refer to an object. Some of the phrases were preceded by a filled pause, and others had no preceding filled pauses. We predicted that listeners would be able to process a complex phrase more quickly when there is a filled pause before the phrase than when there is no filled pause. We assumed that filled pauses give a cue to the speaker's planning difficulty and allow listeners some time to predict the content of the upcoming speech. Consequently, listeners will be able to respond to the referent of a complex phrase more quickly with a preceding filled pause than without. On the other hand, as filled pauses are not so frequent before simple phrases as before complex phrases, filled pauses before simple phrases are not likely to help listeners' predictions. As a result, responses to the referent of a simple phrase will not be quicker, or could even be slower with a preceding filled pause than without.

We conducted experiments with Chinese speakers of Japanese as well as with native Japanese speakers to test the hypothesis. We assumed that it is a universal linguistic tendency that filled pauses are more frequent before longer or more complex constituents. However, as disfluencies have language specific aspects in terms of sounds and forms, listeners would probably need to be exposed to spontaneous speech of a non-native language long enough to be able to recognize and process disfluencies in a non-native language effectively. In Japanese, for example, one of the most frequent fillers, *ano* has a function as a demonstrative adjective similar to *that* in English as well as that of a filler. If non-native Japanese speakers only know *ano* as a demonstrative adjective, they may have trouble in processing *ano* as a filler.

It has widely been observed that the effects of experimental conditions depend on the learners' proficiency levels in the language (e.g. Chiang and Dunkel, 1992; Rubin, 1994). Therefore, non-native participants were grouped into three groups according to their proficiency levels in spoken Japanese. The results were compared among the three groups as well as with the results of native Japanese listeners. It was not known what effect filled pauses at phrase boundaries have on non-native listeners because there has hardly been any research directly related to this topic. However, we predicted that the higher their proficiency in spoken Japanese, the more similar their response pattern to that of the native listeners.

2. Experiment 1

2.1. Method

2.1.1. Participants

The participants were thirty university students who were native speakers of Tokyo (standard) Japanese. They were paid 700 Yen each.

2.1.2. Design

A pair of shapes of the same colour were presented on a computer screen, where one consisted of 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 the appearance of the visual stimulus, speech referring to one of the two shapes was played. The participants' task was to press a button corresponding to the shape being referred to as soon as possible. The instructions given to the participants were as follows (translated from Japanese):

> A woman is asking her interlocutor to bring a paper decoration in a certain colour and a shape. Which one is she asking for? Two pieces of paper appear on the computer screen. Please press either the left or right mouse button corresponding to the piece of paper that she is asking for as soon as possible.

Through these instructions, we aimed at making the participants guess what would follow the filled pauses.

In everyday life, it is not uncommon to infer what the speaker is going to say next when the speaker pauses or hesitates, especially when one is interested in the talk or when one wants to be cooperative with the speaker. We assumed that the participants would do what they normally do even in an experimental setting, although the utterances of this experiment are not ones that would typically be heard repeatedly in everyday life. The answer choices were limited to two in the experiment and this may seem unnatural at a glance. However, it is not unusual in everyday life that the possible following types of utterances are limited by the situational or linguistic contexts and highly predictable for the listener. For example, if a family member says at the breakfast table, "Can you pass me theee um . . .", the possible word following *um* will be a noun or an adjective and the referent of the noun phrase will be limited to one of the items on the table or in the space which the speaker is looking at. The types of utterances in the second parts of adjacency pairs, such as answers to yes-no questions, responses to requests or invitations, are also highly predictable.

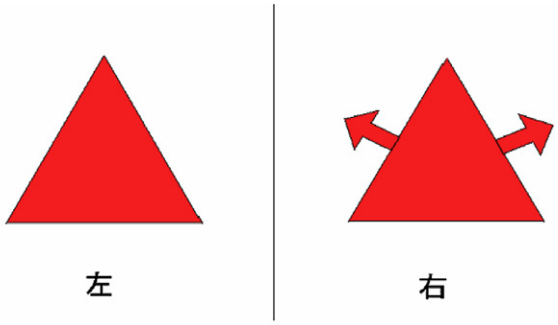

Fig. 1. An example of visual stimuli. The visual stimuli always consisted of a simple shape (round, square or triangular) and a compound shape (with two arrows attached to the simple shape) in the same colour. Chinese characters were allocated under the figures indicating "left" for the left figure and "right" for the right figure.

Each utterance contained a phrase describing a colour (called "a colour phrase") and a phrase describing a shape ("a shape phrase") in this order. As paired shapes were always in the same colour, the colour phrases were not relevant to the task. The target phrases were shape phrases. The experiment involved two factors:

(1) Complexity factor: the speech stimuli referred to either a simple shape or a compound shape ("simple condition" and "complex condition", respectively). We assumed one is aware that more words or a complex structure is usually necessary to describe compound shapes than simple shapes, and therefore, it takes one longer to plan constituents referring to compound shapes than those of simple shapes.
(2) Fluency factor: Target phrases (i.e. shape phrases) were immediately preceded by the filled pause *eto* pronounced most frequently /ɛ:to/ (the vowels and the consonant of which can be shorter or longer), a silent pause of the same duration as a filled pause, or no pause ("filler condition", "pause condition", and "fluent condition", respectively).

Example sentences in simple and complex conditions with filled pauses are given below with English translations. Fillers are in italics. The sentences are shown separated into bunsetu-phrases by spaces. Bunsetu-phrases are linguistic units comprising one content word with or without function words. One or more bunsetu-phrases compose a prosodic phrase in fluent speech.

(1) Simple condition with a filled pause:
Ano-ne, watasi-no heya-kara akaku-te *eto* marui
Dem-Par I-Gen room-Abl red-Con *um* circular
kami mottekite-kureru?
paper bring-Aux
Look, could you bring red and *um* circular paper from my room?
(2) Complex condition with a filled pause:
Ano-ne, watasi-no heya-kara akaku-te *eto* maru-ni
Dem-Par I-Gen room-Abl red-Con *um* circle-Dat
yajirusi-ga tui-ta kami mottekite-kureru?
arrow-Nom attach-Aux paper bring-Aux
Look, could you bring red and *um* circular paper with arrows from my room?
(Abl: ablative; Aux: auxiliary; Con: connective, Dat: dative; Dem: demonstrative; Gen: genitive; Nom: nominative; Par: particle).

In the simple condition, shape phrases were composed of one bunsetu-phrase with one or two words. In the complex condition, shape phrases consisted of three bunsetu-phrases, each of which contained two words. Shape phrases in the complex condition were both longer and more complex than those in the simple condition.

The filler *eto* was used because it is one of the most frequent types of fillers both in dialogues and monologues.

This sound form exclusively functions as a filler, while most other types have functions other than fillers. Another reason was that *eto* was chosen by eight out of ten native speakers of Japanese who were asked to choose one filler they would use in this context among the four frequent types, *ano, e*, *eto,* and *ma*. All of the informants reported that *ano* and *e* could also be used in this context, but not *ma*. Functions of *eto, ano*, and *e*, in this context, therefore, seem very similar, and it is likely that findings of the experiments with *eto* can be generalised to *ano* and *e*.

The silent pause condition was added for the following reasons. When the results in the filler condition differed from those in the fluent condition, we would not know whether the difference was attributable to the time that filled pauses allowed listeners or both the time and the sound of them. Comparison of the effects of filled pauses with those of silent pauses of the same duration would reveal whether the effects are derived from the time or both the time and the sound of filled pauses. If the results of the filler condition differed from those of the pause condition, the difference would be attributable to the sounds of fillers. If not, it is likely that the effects of filled pauses mainly resulted from the time they took rather than their sounds.

2.1.3. Auditory and visual stimuli

Speech stimuli were created in the following way. The first author uttered sentences with a structure as follows:

Ano-ne, $1-kara $2-te *eto* $3 kami
Dem-Par place-from colour-and *um* shape paper

mottekite-kureru?
bring-Aux

Each slot was filled with one of the words or phrases in Japanese below. The list in Japanese is given in Appendix.

$1: top of the desk, my room, next room
$2: black, blue, brown, green, grey, orange, pink, purple, red, yellow
$3: circular, circle with arrows, square, square with arrows, triangular, triangle with arrows

Thus, 180 sentences were created.
$1 (3 items) × $2 (10 items) × $3 (6 items) = 180

Although the test stimuli were presented to the speaker as a reading list, the speaker uttered the sentences without looking at the list so that the utterances sounded like natural, everyday speech. The utterances were recorded in an acoustically treated recording studio. The speech was sampled at 44 kHz and digitised at 16 bits directly onto a PC.

All the utterances contained the filled pause *eto* between the colour phrase and the shape phrase. All the *eto* had a short silent pause before *e*, which is in accordance with the tendency that the majority of *eto* are preceded by a silent pause (Watanabe et al., 2000). Most *eto* also had a short silent pause after *o*. The onsets of shape phrases were one of the following consonants: /m/ for *maru* (circle); /s/ for *sankaku* (triangle); /c/ (voiceless alveolo-palatal median laminal fricative) for *sikaku* (square).

We called the original speech *a filler version*. The original speech was digitally edited and two new versions were created. Fig. 2 shows the editing process of the original speech.

(1) *A pause version*: filled pauses were substituted by a silent period of the same duration.
(2) *A fluent version*: filled pauses were edited out.

Short silences on both sides of the voiced parts of *eto* enabled us to modify speech without making editing noises or creating unnatural F0 contours.

Speech stimuli were checked by two native speakers of Japanese. They reported no unnaturalness about the speech. Table 1 shows the mean duration of speech stimuli in each condition and the mean durations of silent and filled pauses in simple and complex conditions. In the fluent condition, sentences with complex target phrases were 993 ms longer than sentences with simple target phrases (4912 ms vs. 3919 ms), as the former contained eight moras more than the latter. Filled pauses in the complex condition were 224 ms longer than those in the simple condition (953 ms vs. 729 ms).

Three sets of stimuli (A, B, C), each of which contained 180 sentences, were created so that only one of the three versions of the same utterance appeared in each stimuli set. The amplitude of speech stimuli was normalised.

Visual stimuli were created so that half of the correct answers were assigned to the left mouse button and the other half to the right mouse button in each condition in each stimuli set. The experiment was set up using SuperLab Pro.

2.1.4. Procedure

The experiment was individually carried out in quiet rooms at Chiba University and the University of Tokyo 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 regardless of when the participants pressed the response button. Time out was set at about 300 ms after the end of sentences. When participants pressed a button more than once, only the first answer was taken. There were three second intervals between the trials. The experiment lasted 35 min excluding the practice session and a short break in the middle.

After having finished 180 trials, all the participants were asked whether they found anything unnatural in the speech which they heard. None of the participants reported any unnaturalness.

Response times from the beginning of the sound files were automatically measured. The onset of the first words describing a shape was manually marked referring to speech sound, sound waves and sound spectrograms. In the example sentences (1) and (2), the word onsets were marked at

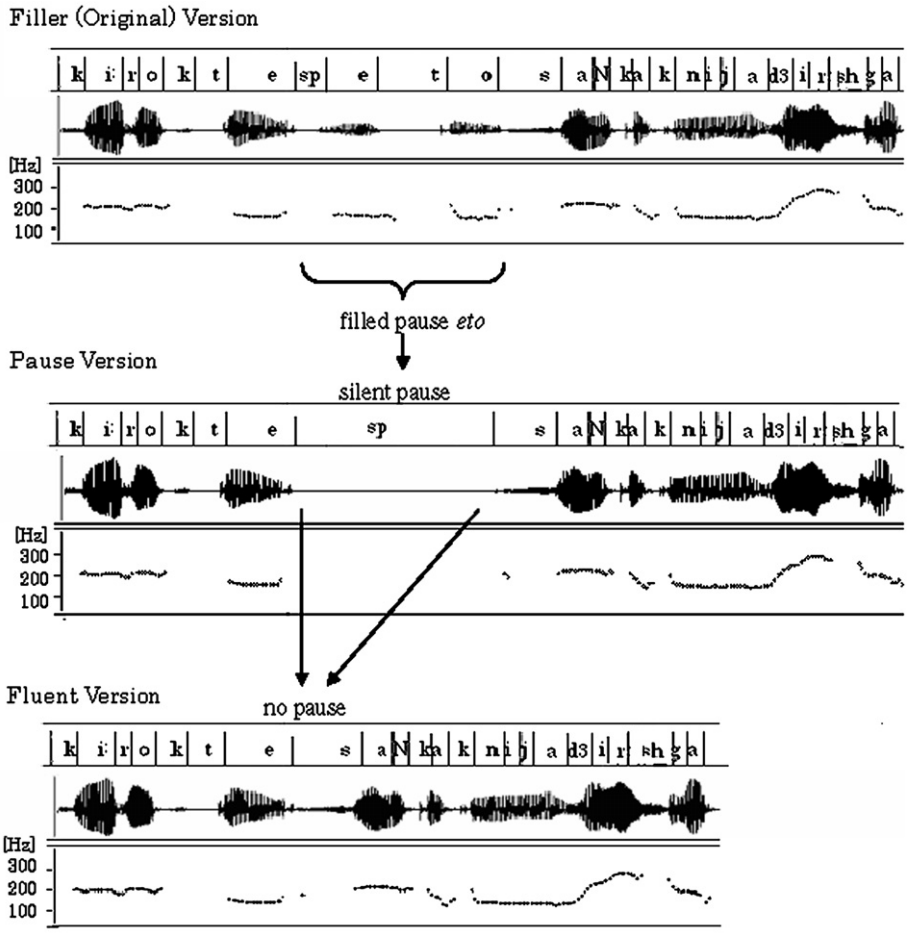

Fig. 2. An example of editing of speech stimuli, with their transcriptions, speech waves, and F0 contours. *Sp* in the transcription stands for *silent pause*. The filler (original) version contained *eto* between "kiirokute" (yellow and) and "sankaku" (triangle). The pause version was created by substituting *eto* with silence of the same duration. The fluent version was produced by editing out *eto* with the adjacent pauses (a preceding pause in this sample) from the filler version.

Table 1
Mean duration of speech stimuli in each condition and mean durations of filled pauses in simple and complex conditions

Conditions	Mean durations (ms)
Simple	
Fluent	3919
Filler/pause	4648
(Filled pauses)	729
Complex	
Fluent	4912
Filler/pause	5865
(Filled pauses)	953

the beginning of /m/ in *maru* (circle). 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 median of the correct response times in each condition for each participant was taken, and the mean median times across conditions were compared.

2.2. Results

One sentence was excluded from analysis because of a defect of the experiment. Three sentences, the error rates of which exceeded 20% in one of the stimuli sets, were also excluded from analysis. The participants pressed a mouse button roughly 500–550 ms after the onset of disambiguating moras. In case of the example sentences 1 and 2 in Section 2.1.2, *i* and *ni* after *maru* (circle) are disambiguating moras. The response button was typically pressed when the fifth mora from the disambiguating mora was being pronounced. It was rare that the button was pressed before the disambiguating mora was heard. Responses made before the onset of disambiguating moras were treated as errors. The mean correct response rate was 98.3%. The rate of correct responses in each condition is given in Table 2. Error responses were excluded from the subsequent analysis.

The mean response times from the shape word onset in the six conditions are shown in Fig. 3. A two-way repeated measures analysis of variance (ANOVA) revealed a main effect of the complexity factor, $F(1,29) = 74.66$, $p < .001$, but no main effect of the fluency factor, $F(2,58) = 1.40$, $p = .25$. A complexity-fluency interaction was significant, $F(2,58) = 5.85$, $p < .005$. Post-hoc tests showed a significant difference among fluency conditions in the complex condition, $F(2,28) = 6.31$, $p < .005$, but no significant

Table 2
Correct response rate of Japanese participants in each condition

Conditions	Rates of correct responses (%)
Simple	
Fluent	98.1
Filler	98.9
Pause	98.9
Complex	
Fluent	97.6
Filler	99.2
Pause	97.0

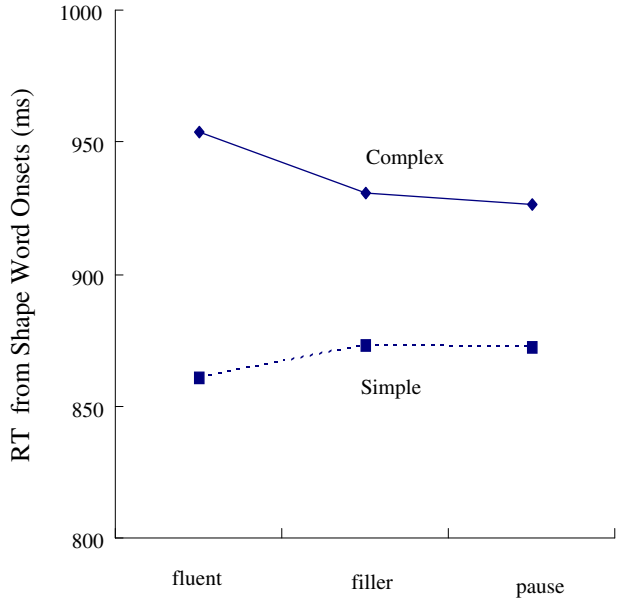

Fig. 3. Japanese participants' mean response times (RT) from the shape word onsets. RT to complex phrases in the filler and the pause conditions were significantly shorter than RT in the fluent condition. There was no significant difference among fluency conditions in RT to simple phrases.

difference in the simple condition, $F(2,28) = 1.20$, $p = .32$. Paired comparisons (alpha adjusted by Bonferroni) revealed that response times to complex phrases in the filler condition and in the pause condition were significantly shorter than those in the fluent condition, $t(29) = 3.13$, $p < .012$; $t(29) = 3.16$, $p < .011$, respectively. There was no significant difference between filler-pause conditions, $t(29) = 0.49$, $p = 1.00$.

Two-way analyses of variance (ANOVA) over items in three stimuli sets (A, B, C) all showed main effects of the complexity factor, A: $F(1,170) = 36.46$, $p < .001$; B: $F(1,170) = 17.74$, $p < .001$; C: $F(1,170) = 52.71$, $p < .001$, but no main effects of the fluency factor, A: $F(2,170) = .26$, $p = .77$; B: $F(2,170) = .81$, $p = .45$; C: $F(2,170) = 1.71$, $p = .18$. A complexity-fluency interaction was not significant in any of the stimuli sets, A: $F(2,170) = .65$, $p = .53$; B: $F(2,170) = .53$, $p = .59$; C: $F(2,170) = .50$, $p = .61$.

2.3. Discussion

Response times to complex phrases were shorter when the phrases were immediately preceded by a filled pause than when there was no pause. On the other hand, no significant difference was found with simple phrases between filler-fluent conditions. These findings suggest that only the filled pauses before complex phrases facilitate listeners' processing of the upcoming speech. These results agree with our prediction that filled pauses cause listeners to expect referents which are likely to be expressed by a longer or more complex constituent.

It is possible that the different response patterns in simple and complex conditions are partly due to difference in the pause duration. Filled pauses in the complex condition were 224 ms longer than those in the simple condition on average (953 ms vs. 729 ms). However, this difference is irrelevant to the observed effects of filled pauses in the complex condition because the effects were compared between filler-fluent conditions in the complex condition.

The prosody of phrases before shape phrases cannot have influenced the effects of fluency conditions because fluent, filler and pause versions of each utterance had exactly the same speech signal except for the part of filled pauses.

One may argue that the responses to compound shapes were quicker with filled pauses than without because filled pauses allowed the listeners extra time for processing the foregoing speech and better prepared them for the following speech, not because filled pauses triggered certain type of expectations. However, this is not likely. Firstly, responses to simple phrases as well as complex phrases should have been faster after filled pauses if this had been the case. Responses to simple phrases were not quicker with filled pauses than without, despite the fact that the phrases up to the shape words had the same semantic contents in both conditions. Secondly, it is unlikely that the listeners needed a pause to process the foregoing colour words. Two shapes of the same colour were on the computer screen when they heard the colour words, and the information about the colour was conveyed early in the phrase. In the case of *akaku-te* (red and), for example, the part which listeners need to know the colour is the first two moras, *aka,* because *ku* is an inflexional suffix and *te* is a connective. Therefore, the colour must have been recognized by the listeners before the end of the phrase. For these reasons, it is unlikely that the native listeners needed the following pause to process the colour words.

It is possible that listeners' responses to compound shapes were quicker in the filler condition than in the fluent condition because they expected the more unfamiliar object after filled pauses. Our hypothesis still holds in this case because one generally needs more words to describe unfamiliar objects than familiar items. Response times to simple shapes after filled pauses were not delayed possibly because shape phrases in the simple condition were easy enough for native listeners to process despite incorrect expectations.

There was no significant difference between filler-pause conditions in responses to the complex phrases. This result suggests that the time which passes during the pauses is critical for the effects. Although no difference was observed between filler-pause conditions, there is a need to further examine whether the two types of pauses have the same effects in other syntactic contexts. Speakers pause at high probabilities at deep syntactic boundaries such as sentence and clause boundaries. Such pauses divide speech stream into meaningful units and give listeners time for processing (Sugito, 1990). Therefore, silent pauses at deep syntactic boundaries are unlikely to be perceived as disfluencies, even if parts of the pauses are used for speech planning, unless they are extremely long. The effects of filled pauses at such locations may well differ from those of silent pauses, because filled pauses are likely to signal speakers' planning difficulties more explicitly (Fox Tree, 2002).

In the next section we test the same hypothesis that listeners expect a relatively long or complex phrase when there is a filled pause with native Chinese speakers of Japanese.

3. Experiment 2

3.1. Method

3.1.1. Participants

Forty-four native speakers of Chinese who had been staying in Japan for more than half a year and had a fairly good command of everyday Japanese took part in the experiment in exchange for 700 Yen. All the participants were either students or researchers at Chiba University or the University of Tokyo, in Japan. As no records of the participants' proficiency levels in Japanese were available, we employed the length of stay in Japan as an index of their proficiency in comprehension of spoken Japanese. We recruited participants so that one third would be those staying in Japan for a period of 0.5–1.5 years (novice group), another one third for 1.5–2.5 years (intermediate group), and the other one third longer than 2.5 years (expert group). We did not include Chinese speakers staying in Japan for less than half a year because their ability to comprehend spoken Japanese could be too divergent to be classified in one group. We also intended to avoid recruiting participants whose listening proficiency was not high enough for the task. The time interval of length of stay for grouping participants was based on our observation that international students using Japanese in their everyday lives make great strides in oral proficiency in the first year after arrival and hardly have any problem carrying on everyday conversation approximately in the third year.

Three participants were excluded from analysis because they turned out to be bilingual speakers of Chinese and other languages. We also excluded participants from analysis if the number of error trials, including trials which timed out, exceeded 10% of the presented trials to avoid too many missing values. They were substituted with newly recruited participants who belonged to the same proficiency groups. Five participants were substituted for this reason. Thus, data from 36 participants, 12 in each proficiency group, were analysed.

3.1.2. Design

The design was the same as Experiment 1 except that the factor of listeners' proficiency in Japanese was added to the analysis. The proficiency factor had three conditions according to the participants' length of stay in Japan: "novice" (0.5–1.5 years), "intermediate" (1.5–2.5 years), and "expert" (longer than 2.5 years).

3.1.3. Procedure

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

After having finished 180 trials, all the participants were asked whether they found anything unnatural in the speech which they heard. None of the participants reported any unnaturalness.

3.2. Results

One sentence was excluded from analysis because of a defect of the experiment. There was no sentence of which the error rate exceeding 20% in any one of the stimuli sets. It was rare that the button was pressed before the disambiguating mora was heard. Responses made before the onset of disambiguating moras were treated as errors. The mean correct response rate was 97.5%. The rate of correct responses in each condition is given in Table 3. Error responses were excluded from the subsequent analysis. The median of the correct response times from the word onset in each condition for each participant was calculated and the mean median times across conditions were compared.

The mean response time from the onset of shape words in each condition for each proficiency group is shown in Fig. 4. The results of a three way analysis of variance (ANOVA)

Table 3
Correct response rate of Chinese participants in each condition (%)

Conditions	Novice	Intermediate	Expert
Simple			
Fluent	98.6	96.1	98.6
Filler	97.2	96.3	94.9
Pause	96.6	98.9	97.5
Complex			
Fluent	97.8	98.3	96.7
Filler	98.3	98.1	97.8
Pause	97.8	97.8	98.3

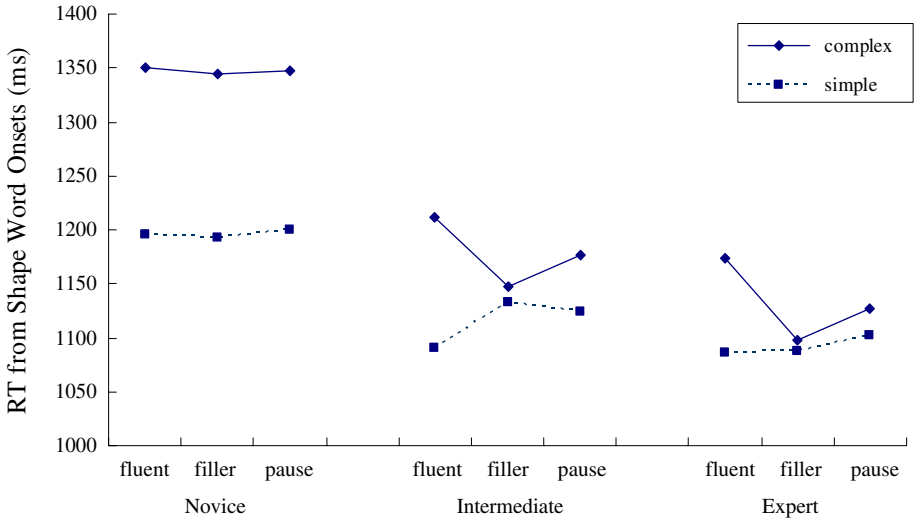

Fig. 4. Chinese groups' mean response times (RT) from the shape word onsets. In the novice group (left), only a main effect of the complexity factor was significant. In the intermediate group (middle), RT to complex phrases in the filler condition were significantly shorter than RT in the fluent condition. In contrast, RT to simple phrases in the filler condition were significantly longer than RT in the fluent condition. There was no significant difference between fluent-pause or filler-pause conditions either in the complex or simple conditions. In the expert group (right), RT to complex phrases in the filler and the pause conditions were significantly shorter than those in the fluent condition. There was no significant difference in RT to simple phrases among any fluency conditions.

with two repeated-measure factors (the complexity factor and the fluency factor) and one between-groups factor (the proficiency factor) are given in Table 4. There were significant main effects of the complex factor and the fluency factor, $F(1,33) = 12.95$, $p < .001$; $F(2,66) = 4.88$, $p < .011$, respectively. Interactions between complexity-fluency factors and complexity-fluency-proficiency factors were also significant, $F(2,66) = 11.20$, $p < .001$; $F(4,66) = 2.51$, $p < .05$, respectively. We inspected the three way interaction by proficiency factor.

In the novice group, there was a simple main effect of the complexity factor, $F(1,33) = 13.76$, $p < .001$, but no simple main effect of the fluency factor, $F(2,66) = .16$, $p = .85$. A simple interaction between complexity-fluency factors was not significant, $F(2,66) = .063$, $p = .939$. No significant difference among fluency conditions was found either in the simple or the complex condition, $F(2,132) = .104$, $p = .901$; $F(2,132) = .103$, $p = .903$, respectively. Response times to complex phrases were significantly longer than response times to simple phrases in all the fluency conditions.

In the intermediate group, there was no simple main effect of the complexity factor or the fluency factor, $F(1,33) = 2.35$, $p = .13$; $F(2,66) = .69$, $p = .51$, respectively. A simple interaction between complexity-fluency factors was significant, $F(2,66) = 10.23$, $p < .001$. Post-hoc tests revealed that there were significant differences among fluency conditions both in the simple and the complex conditions, $F(2,132) = 4.07$, $p < .019$; $F(2,132) = 8.62$, $p < .001$, respectively. Paired comparisons (alpha adjusted by Bonferroni) showed that response times to complex phrases were significantly shorter in the filler condition than in the fluent condition, $t(11) = 3.82$, $p < .002$. There was no significant difference between fluent-pause conditions or filler-pause conditions, $t(11) = 2.51$, $p = .052$; $t(11) = 1.73$, $p = .28$, respectively. In contrast, response times to simple phrases were significantly longer in the filler condition than in the fluent condition, $t(11) = 2.65$, $p < .037$, but there was no significant difference between fluent-pause conditions or filler-pause conditions, $t(11) = 2.11$, $p = .13$; $t(11) = 0.69$, $p = 1.00$, respectively.

In the expert group there was a significant simple main effect of the fluency factor, $F(2,66) = 6.94$, $p < .002$, but no simple main effect of the complexity factor, $F(1,33) = .98$, $p = .33$. A simple interaction between complexity-flu-

Table 4
Results of the analysis of variance

Source	df	MSe	F	p
Between subjects				
Proficiency (P)	2	505095.63	3.15	.056
Error (P)	33	160139.31		
Within subjects				
Complexity (C)	1	385810.14	12.95	.001**
$C \times P$	2	61724.91	2.07	.142
Error C (P)	33	29800.47		
Fluency (F)	2	5758.75	4.88	.011*
$F \times P$	4	1716.35	1.46	.226
Error F (P)	66	1179.04		
$C \times F$	2	19169.35	11.20	.001**
$C \times F \times P$	4	4304.89	2.51	.050*
Error CF (P)	66	1712.17		

* $p < .05$.
** $p < .01$.

ency factors was significant, $F(2,66) = 5.93$, $p < .004$. Post-hoc tests revealed that there was a significant difference among fluency conditions in the complex condition, but not in the simple condition, $F(2,132) = 12.00$, $p < .001$; $F(2,132) = .64$, $p = .53$, respectively. Paired comparisons (alpha adjusted by Bonferroni) showed that response times to complex phrases were significantly shorter in the filler condition than in the fluent condition, $t(11) = 4.49$, $p < .001$. Response times to complex phrases in the pause condition were also significantly shorter than those in the fluent condition, $t(11) = 3.18$, $p < .01$. There was no significant difference between filler-pause conditions: $t(11) = 1.82$, $p = .23$.

3.3. Discussion

Chinese groups showed different response patterns depending on their proficiency in Japanese. The response pattern of the expert group was parallel to that of native Japanese speakers, although the mean response time was 209 ms longer than that of native listeners. Namely, the response times to complex phrases were shorter with preceding filled pauses than without any pauses, whereas no significant difference was found in response times to simple phrases between any fluency conditions. Response times to complex phrases in the silent pause condition were also shorter than those in the fluent condition, and there was no significant difference between filler-pause conditions. The results of the expert group agreed with our prediction mentioned in Section 1.4 and supported the hypothesis. The expert group seemed to be utilising filled pauses at phrase boundaries just like native Japanese listeners.

In contrast with the expert group, the response pattern of the novice group was different from that of the Japanese listeners. Response times to complex phrases were simply longer than those to simple phrases in all the fluency conditions. No effects of filled or silent pauses were observed either in the simple or the complex condition. The hypothesis was not supported with this group. The novice group might have been unaware of the occurrence of filled pauses, or they were aware of them but could not utilize them as native or expert non-native listeners do because of their limited language proficiency.

The response pattern of the intermediate group is interesting in that it seems to reveal a transition stage of learners acquiring native-like strategies for processing filled pauses. Their response pattern to complex phrases is similar to those of expert and native listeners. Their responses to complex phrases were quicker with preceding filled pauses than without any pause. This is consistent with the results of native and expert listeners. On the other hand, the response times to simple phrases in the filler condition were significantly longer than those in the fluent condition. This result indicates that filled pauses before a simple phrase lead listeners to an incorrect expectation. This response pattern can happen when listeners have expected a compound shape after filled pauses and their expectation is betrayed. The delayed response to simple phrases can be a result of their limited language proficiency. If the listeners' proficiency had been high enough to cope with an unexpected phrase or to be aware that a simple phrase may also follow filled pauses, the observable delay could have been avoided. We infer that this is the case with native and expert non-native groups: Their responses to simple phrases after filled pauses were not delayed because of their high language proficiency.

In the intermediate group, there was no significant difference between fluent-pause or filler-pause conditions either in complex or simple conditions, although the difference between fluent-pause conditions in the complex condition tended to be significant. The difference between filler-fluent conditions was clearer than the difference between pause-fluent conditions both in complex and simple conditions, although there was no significant difference between filler-pause conditions. These results support the view that filled pauses signal speakers' planning difficulties more explicitly than silent pauses of the same duration.

4. General discussion

Native listeners' responses to compound shapes were quicker when there was a filled pause before the phrase describing the shape than without any pause. On the other hand, no significant difference was observed in response times to simple shapes between fluent-filler conditions. These results support our prediction that filled pauses bias listeners' expectations toward referents which tend to be expressed by a longer or more complex constituent.

Our findings support the argument of Bailey and Ferreira (2003) that listeners are sensitive to the co-occurrence of filled pauses and the onset of a larger constituent. Our results agree with those of Arnold et al. (2003) in that disfluencies cause listeners to expect that the speaker is going to refer to items which will be more difficult to express and take him or her longer to plan. Arnold, Fagnano, and Tanenhaus focused on the difficulty in word access in terms of discourse status of the referent, whereas we considered the difficulty from the viewpoint of syntactic complexity.

Various factors can be involved with the length and complexity of constituents expressing objects, which leads to the rate of the preceding disflencies. In this study we assumed that the number of component objects was relevant to the length and complexity of phrases referring to the items. It could also be presupposed that the degree of familiarity with objects or the amount of information required to specify items is related to constituent length and complexity. Unfamiliar objects or items surrounded by similar articles tend to be expressed by longer or more complex constituents than familiar or unique items. Therefore, constituents referring to the former type of objects are more likely to be preceded by disfluencies than the latter, and listeners can be sensitive to the types of objects following disfluencies.

As our study was limited to the effects of one type of filler *eto*, we need to examine the effects of other types to

understand the functions of filled pauses more comprehensively. However, as *ano* and *e*, which are among the four frequent types of fillers, are also natural in this context, experiments using *ano* or *e* instead of *eto* are likely to result in similar findings.

We conjecture that filled pauses, repetitions and prolongations of function words, and the combination of those such as *theee um* and *and uh* in English can have similar effects to those of *eto* in this study. In Japanese, articles do not exist and function words are located after content words in constituents (e.g. postpositions instead of prepositions). Therefore, less function words are available as a time gaining device at the beginning of major constituents (Fox et al., 1996). Thus, filled pauses seem to play a more dominant role among other types of disfluencies in Japanese than in English.

In the present study, we examined the effects of filled pauses at bunsetu-phrase boundaries. It remains to be seen what influence filled pauses within bunsetu-phrases have on listeners. When the sounds of fillers such as *ano* and *ma* are not delimited by silent pauses and compose a part of prosodic phrases, it may be more difficult for listeners, particularly for non-native speakers, to recognize and process them as fillers.

Chinese speakers' response patterns varied depending on their proficiency in spoken Japanese. Filled pauses had neither negative nor positive effects on those who had stayed in Japan for 0.5–1.5 years. However, in 1.5–2.5 years, Chinese speakers studying at Japanese universities seem to become aware of the forms and functions of filled pauses in Japanese and start utilizing them in communication in a way similar to that of native Japanese listeners. Those who have stayed in Japan for longer than 2.5 years seem to have the same strategies for processing filled pauses as those of Japanese listeners.

The results for Chinese speakers did not show any negative effects of filled pauses at phrase boundaries either, except for the case of the second year group: Filled pauses before simple phrases delayed their responses. However, it is unlikely that the delay was caused because the listeners of this group did not recognize fillers as such, as Buck (2001) argued. Positive effects of filled pauses before complex phrases would not have been observed if the listeners had not recognized fillers as fillers. These negative effects seem to reflect a stage of Chinese listeners acquiring the usage of filled pauses in the non-native language.

In examining the effects of filled pauses on non-native listeners, a finer control of listeners' proficiency levels in the non-native language and the lengths of stay, particularly in the first two years after arrival, will be useful to obtain more fine-grained pictures of listeners acquiring skills of processing filled pauses in the non-native language.

5. Conclusion

We tested the hypothesis that listeners expect a relatively long or complex phrase when there is a filled pause. Japanese speakers' response times to compound shapes were shorter when the phrases describing the shape were preceded by a filled pause than when there was no pause. On the other hand, no significant difference was observed in response times to simple shapes between fluent-filler conditions. These results provided evidence to support our hypothesis. Our findings together with the results of previous studies indicate that filled pauses cause listeners to expect that the speaker is going to refer to something that will take him or her longer to plan.

The results with Chinese speakers shed light on the stages of non-native listeners acquiring skills for processing filled pauses in a non-native language. Further research on disfluency processing by non-native listeners will surely be useful not only for studies of foreign language learning and teaching but also for understanding language-universal and language-specific aspects of human speech processing.

Acknowledgements

We thank Max Coltheart and Sallyanne Palethorpe for their invaluable advice in planning the experiments. This work was partly supported by JST/CREST through the Expressive Speech Processing Project (2001–2005).

References

Arnold, J.E., Wasow, T., Losongco, T., Ginstrom, R., 2000. Heaviness vs. Newness: the effects of structural complexity and discourse status on constituent ordering. Language 76 (1), 28–55.

Arnold, J.E., Fagnano, M., Tanenhaus, M.K., 2003. Disfluencies signal theee, um, new information. J. Psycholinguist. Res. 32 (1), 25–36.

Bailey, K.G.D., Ferreira, F., 2003. Disfluencies affect the parsing of garden-path sentences. J. Mem. Lang. 49, 183–200.

Blau, E.K., 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).

Brennan, S.E., Schober, M., 2001. How listeners compensate for disfluencies in spontaneous speech. J. Mem. Lang. 44, 274–296.

Brennan, S.E., Williams, M., 1995. The feeling of another's knowing: prosody and filled pauses as cues to listeners about the metacognitive states of speakers. J. Mem. Lang. 34, 383–398.

Buck, G., 2001. Assessing Listening. Cambridge University Press, Cambridge.

Chiang, C.S., Dunkel, P., 1992. The effect of speech modification, prior knowledge, and listening proficiency on EFL lecture learning. TESOL Quart. 26 (2), 345–374.

Clark, H.H., 2002. Speaking in time. Speech Comm. 36, 5–13.

Clark, H.H., Fox Tree, J.E., 2002. Using uh and um in spontaneous speaking. Cognition 84, 73–111.

Clark, H.H., Wasow, T., 1998. Repeating words in spontaneous speech. Cognitive Psychol. 37, 201–242.

Cook, M., Smith, J., Lalljee, M.G., 1974. Filled pauses and syntactic complexity. Lang. Speech 17 (1), 11–16.

Finegan, E., 1994, .Second ed. In: Language: Its Structure and Use Harcourt Brace, New York.

Fox, B.A., Hayashi, M., Jasperson, R., 1996. Resource and repair: a cross linguistic study of syntax and repair. In: Ochs, Schegloff, E.A.,

Thompson (Eds.), 2004a In: Interaction and Grammar, vol. 185–237. Cambridge University Press, Cambridge.
Fox Tree, J.E., 1995. The effects of false starts and repetitions on the processing of subsequent words in spontaneous speech. J. Mem. Lang. 34, 709–738.
Fox Tree, J.E., 2001. Listeners' uses of um and uh in speech comprehension. Mem. Cognition 29 (2), 320–326.
Fox Tree, J.E., 2002. Interpreting pauses and ums at turn exchanges. Discourse Process. 34 (1), 37–55.
Fukao, Y., Mizuta, S., Ohtsubo, K., 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).
Goldman-Eisler, F., 1968. Psycholinguistics. Academic Press, London.
Griffiths, R., 1991. Pausological research in an L2 context: a rationale, and review of selected studies. Appl. Linguist. 12 (4), 345–364.
Holms, V.M., 1988. Hesitations and sentence planning. Lang. Cognitive process. 3 (4), 323–361.
Levelt, W.J.M., 1989. Speaking. The MIT Press, Cambridge, Massachusetts.
Maclay, H., Osgood, C.E., 1959. Hesitation phenomena in spontaneous English speech. Word 15, 19–44.
Rose, R.L., 1998. The Communicative Value of Filled Pauses in Spontaneous Speech, Unpublished MA thesis submitted to University of Birmingham.
Rubin, J., 1994. A Review of Second Language Listening Comprehension Research. Mod. Lang. J. 78, 199–221.
Sadanobu, T., Takubo, Y., 1995. The monitoring devices of mental operations in discourse – a case of 'eeto' and 'ano (o)' –. Gengo Kenkyu 108, 74–93, in Japanese.
Shriberg, E.E., 1994. Preliminaries to a theory of speech disfluencies. Unpublished Ph.D. thesis, University of California at Berkeley.
Shriberg, E.E., 2005. Spontaneous speech: How people really talk, and why engineers should care. In: Proc. 9th European Conf. on Speech Comm. and Technol., Lisbon, Portugal, pp. 1781–1784.
Stenstroem, A., 1994. An Introduction to Spoken Interaction. Longman, London and New York.
Sugito, M., 1990. On the role of pauses in production and perception of discourse. In: Proc. 1st Internat. Conf. Spoken Lang. Process., Kobe, Japan, pp. 513–516.
The National Institute for Japanese Language, 2004. The Corpus of Spontaneous Japanese homepage. <http://www2.kokken.go.jp/%7Ecsj/public/6_1.html>, (retrieved in April, 2006).
Voss, B., 1979. Hesitation phenomena as sources of perceptual errors for non-native speakers. Lang. Speech 22 (2), 129–144.
Wasow, T., 2002. Postverbal Behavior. CSLI Publications, Stanford, California.
Watanabe, M., 2003. The constituent complexity and types of fillers in Japanese. In: Proc. 15th Internat. Congress of Phonetic Sci., Barcelona, Spain, pp. 2473–2476.
Watanabe, M., Ishii, C.T., 2000. The distribution of fillers in lectures in the Japanese language. In: Proc. 6th Internat. Conf. on Spoken Lang. Process., vol. 3, Beijing, China, pp. 167–170.
Watanabe, M., Den, Y., Hirose K., Minematsu, N., 2004a. Types of clause boundaries and the frequencies of filled pauses. In: Proc. 18th Annual Convention of the Phonetic Society of Japan, (in Japanese) pp. 65–70.
Watanabe, M., Den, Y., Hirose, K., Minematsu, N., 2004b. Clause types and filled pauses in Japanese spontaneous monologues. In: Proc. 8th Internat. Conf. Spoken Lang. Process., Jeju Island, Korea, pp. 905–908.

Appendix

List of speech stimuli

Structure

Ano-ne	$1-kara	$2-te (de)	*eto*	$3	kami	mottekite-kureru?
Dem-Par	place-Abl	colour- Con	*um*	shape	paper	bring-Aux

$1: locations

tukue-no ue	tonari-no heya	watasi-no heya
desk-Gen top	next room	my room

$2: colours

kuro-ku	ao-ku	chairo-ku	midori	haiiro	orenzi	pinku
black	blue	brown	green	grey	orange	pink

murasaki	aka-ku	kiiro-ku
purple	red	yellow

$3: shapes

maru-i	maru-ni yazirusi-ga tui-ta
circular	circle-Dat arrow-Nom attach-Aux

sikaku-i	sikaku-ni yazirusi-ga tui-ta
square	square-Dat arrow-Nom attach-Aux

sankaku-no	sankaku-ni yazirusi-ga tui-ta
triangle-Gen	triangle-Dat arrow-Nom attach-Aux

Abbreviations:

Abl	ablative
Aux	auxiliary
Con	connective
Dat	dative
Dem	demonstrative
Gen	genitive
Nom	nominative
Par	particle