Towards a Grammar of Spoken Language
- Prosody of Ill-formed Utterances and Listener’s Understanding in Discourse -

Miyoko Sugito
Institute for Speech Communication Research, Nara Japan
msugito@nifty.ne.jp

Abstract
So-called ill-formed utterances abound in natural discourse. This paper describes how listeners can understand such utterances in a discourse, including speech hesitations and slips-of-the-tongue. Through analysis of durations of utterances and their succeeding pauses, and of intonation patterns, and also through listener understanding experiments using ill-formed utterances, the following results are obtained.
1) Listeners can understand and memorize speech contents by compensating for the ill-formedness, but pauses are indispensable for the processing.
2) Listeners judge appropriateness of each passage in utterances and pick up significantly important phrases, using intonational peculiarities observed in speech hesitation and truncation.

1. Introduction
Spoken language is no doubt the original form of language, which is apparent from the fact that languages without the written form exist, but not vice versa. However, grammars have been described only for written languages, since sentences of written language are generally well-formed. On the other hand, many utterances of spoken language do not conform with grammar of written language and yet can be easily understood, presumably by the presence of pause and intonation. For example, pause is a duration-based recess in spoken language, but is simply represented by a period or a comma in written language. The intonation, which is expressed by vocal-tone dynamics, cannot be represented in literal text. These prosodic features play very important roles in the production and perception of spoken language. In this paper, we discuss the paradox of how and why ill-formed utterances in discourse can be easily understood by listeners.

2. Characteristics of speech in discourse

2.1. Speech material
The material used here is selected from one of the audio and video recordings of conversations by speakers of Tokyo dialect or Osaka dialect, collected by the research project “Aspects of Text and Context – An Analysis of Conversational Texts” (National Language Research Institute, 1987) in which the author took part [1]. It consists of speech passages from a conversation by four speakers of the Osaka dialect on recollections from their youth. The material is especially rich in natural content since it is the recording of a casual conversation carried on by the four speakers while the moderator temporarily left the room intentionally. The participants consist of a key speaker A and three others (B, C and D). Speaker B occasionally takes over the turn and introduces a new topic, while speaker C often shadows the utterances of A. Speaker D is a good listener and helps speaker A by adequate back-channel responses (henceforth brief responses). The key speaker A is rich in topics and develops his talk naturally. He is a merchant at Semba in Osaka, was wealthy in the old days, and very well remembers the events that he experienced with full curiosity. He talks successively on a number of topics, but his talks are full of ill-formed utterances and speech hesitations (henceforth hesitations). Yet other speakers enjoy joining him in the conversation, and his utterances are easy to understand even for those who listen to the recorded conversation. In order to clarify why one can understand such conversational speech so easily, the prosody of the speech material is examined from the viewpoints of both the speaker and the listener.

2.2. Utterance length and pause length
Here we define an utterance to be a passage of speech between successive pauses. Figure 1 shows the relationship between utterance length and succeeding pause length in speaker A’s speech (total 77.8 sec.). In the figure, ● denote sentence-final utterances, ▲ denote hesitations, and □ denote other utterances. In a cheerful conversation, brief response or laughing voice often overlaps with the speaker’s voice, but the speaker proceeds with his talk uninterrupted. The distributions of utterance lengths and succeeding pause lengths are widely scattered for both ● and □, while utterances and pauses for hesitations are equally brief, so that the corresponding points (▲) are concentrated at the lower left corner of the figure.

Fig. 1 The relationship between utterance length and succeeding pause length in speaker A’s speech.

Since pausing is closely related with inhalation, longer utterances are presumably followed by longer pauses [2]. However, the correlation coefficient between utterance length and pause length is 0.213 for the total data, and is 0.0157 for the data excluding hesitations. These results indicate that the above-mentioned conjecture is not valid. A speaker may inhale rapidly after a long utterance. Thus in a lively conversation, pauses are not controlled merely by the physiological
3. Ill-formed utterances in discourse

3.1. Structure of ill-formed utterances

Shown below is the transcription of the conversational material used in this analysis. The talk is quite natural, but most of the utterances are ill-formed in the sense that they do not follow grammar of written language [4].

Table 1 shows the layered structure of these utterances based on categories of (a) place, (b) time, and (c) event. The first passage (1) "soryaamoo sorya soo danganga" indicates a sudden turn of the topic. The phrase "kinjono" in the passage (1) originally was meant to modify the (9) /"yotsubashino/", and the utterance error (2) /"yott/ indicates that the speaker A was trying to utter "yotsubashi". The phrase "wakaitokinimo" (also in my youth) refers to the 37th year of the Meiji era (1904) in the passage (8). In the written language, these utterances of speaker A can be summarized as follows: "In the 37th year of the Meiji era I went out to Yotsubashi crossing to watch the newly introduced streetcar, together with many others, who all brought mats with them to sit on."

As seen in Table 1, however, places, times and events cross each other in message construction. Such ill-formedness is considered to be the consequence of parallel processing for lexical access, sentence construction, and speech production. Namely, a sudden change in conversation topic may have resulted in time lags between commands for sequential arrangement of the words and for physical movements of the articulators, and have produced ill-formed utterances. However, listeners can usually understand such messages without consciousness of their ill-formedness. As already mentioned above, pause and intonation may be playing key roles for the understanding of such ill-formed utterances.

3.2. Intonation in discourse

Pitch frequency contours and pause locations of the speech material are shown in Fig.2, aligned with the speech signal waveform. Each utterance is numbered, and a brief response, laughing voice and so on are indicated by angle brackets. Arrows in the figure indicate the locations of very low troughs in the pitch frequency contours, where brief responses or interruptions from listeners tend to be inserted.

Table 1 The layered structure of ill-formed utterances of speaker A.

<table>
<thead>
<tr>
<th>&lt;a. place&gt;</th>
<th>&lt;b. time&gt;</th>
<th>&lt;c. event&gt;</th>
</tr>
</thead>
<tbody>
<tr>
<td>In our neighborhood</td>
<td>When young</td>
<td>Street car</td>
</tr>
<tr>
<td>Yotsubashi crossing</td>
<td>Diamond crossing, we called</td>
<td>All with mats to sit</td>
</tr>
<tr>
<td>Yotsubashino koosatende/</td>
<td>All with mats to sit</td>
<td>to go to see</td>
</tr>
</tbody>
</table>

The intonation of the initial section of (1) "soryaamoo/" is lower than that of the immediately following sections of speech, and is characterized by flatness. It is supposed that the interval during which the flatness lasts corresponds to that of information retrieval activities in the brain. Similar intonation is observed in the passage (4) /"anoo/ (well)/, which is also conjectured to correspond to information retrieval activities [5]. In addition to these sections, the latter part of the passage (7) /"sanjuurokunenyakedo/ also shows low and flat characteristics of pitch, in contrast with the normal characteristics of rising and falling tone and accents. The passage (10) /"daiyamondo koosatenyute/ is an embedded sentence, and carries relatively low intonation. The utterance (13) /"hah/ is particularly lowered in pitch, and is slowed down in speed. These features indicate a break in the discourse [6].
3.3. Features of truncation and hesitation

Partial or incomplete utterances are caused by truncation or hesitation. The speech features of these utterances are as follows:

1. The duration is brief, and the succeeding pause is also rather brief.
2. In the intonation,
   (2a) the pitch contour shows unnatural characteristics in the case of truncation,
   (2b) the pitch contour is almost flat in the case of hesitation.

As shown in Fig. 2, the intonation of the passage (2) “yot” belongs to case (2a). When an utterance is discontinued, the intonation pattern is generally normal up to the point at which the utterance is terminated. However, when an utterance is truncated at the beginning as in the passage (2), the pattern is quite peculiar. It is supposed that the interaction between the articulatory command for the initial part of the utterance and the command for cancelling the whole utterance may cause an intonation peculiar to truncation.

On the other hand, the intonation of passage (4) “anoo” (well) maintains low and flat characteristics. This serves as a signal for hesitation, but also indicates that the speaker is trying to keep the turn. As the discourse becomes well-organized, the number of hesitations decreases. Also, the utterance length becomes more uniform, the pauses become shorter, and the ratio between utterance length and pause length becomes less variable. Such low and flat intonation is observed regardless of dialects. Both in the Tokyo dialect which is characterized by a low tone beginning in the accent pattern, and in the Osaka dialect which allows a succession of high tones, hesitations such as “anoo” are uttered in the same manner in spite of the difference of the word accent tone in the two dialects. This is the reason to suggest that the flat intonation might indicate the ongoing process of information retrieval prior to or during an utterance.

3.4. Intonation of utterance and listener’s response

The listener’s brief response or interruption is closely related with the speaker’s intonation. When a speaker hesitates or locates a pause in the course of an utterance, the pitch frequency does not become very low, but at the ends of sentences or semantic groups, it drops significantly. The listener’s brief responses correspond to the declination-tone characteristics of the speaker. This is indicated by the correspondence between falling tones and listener’s responses (shown in angle brackets < >) in Fig. 2. On the other hand, in the final passage (12) “mini ikimanea”, the voices of the speaker and the listeners overlap with each other. This occurs when the speaker and the listeners collaborate in developing the story.

4. Listener’s understanding of ill-formed utterances

4.1. Speech stimuli for the experiments

The following experiments are conducted to clarify how listeners memorize the content of messages which the speaker intends to transmit by ill-formed utterances. The stimulus (1) is prepared from conversational speech data described in the preceding sections, by removing hesitation or embedded speech as well as overlapped or meaningless portions in the conversation.

1. “uchirano kinjoko wakaitokinimo / denshaga meiji / sanjuushichimen / yotsubashino / koosatennde / mina gozamotte suwarini / mini ikimanea”

Grammatical inconsistencies still remain in these reformed stimuli. Based on the assumption that speaker’s pauses help understanding by the listener, fully-connected speech in which all pauses have been removed is used as the stimulus (2) for the experiment.

2. “uchirano kinjoko wakaitokinimo denshaga meiji / sanjuushichimen / yotsubashino / koosatennde / mina gozamotte suwarini mini ikimanea”

Figure 3 illustrates the pitch contour of the speech without brief or redundant speech segments which are not directly related to the development of a meaningful story.

4.2. Experiments

Two experiments were conducted using the above-mentioned speech stimuli (1) and (2). Subjects (as listeners) are fourteen students who live in the Osaka area, seven for experiment-1 and seven others for experiment-2.

[Experiment-1]

Each subject was instructed to listen to the speech stimulus (1) carefully, and to answer the content of the heard speech after making a sequential mental count for fifteen seconds. The task of the mental counting was given to prevent speech content from being transferred to long-term memory by rehearsal during the fifteen seconds of decay period in short-term memory. The total succession of the experiment was recorded on a tape recorder, but the subjects were not informed. Recorded subjects’ answers were transcribed, and the number of recalled phrases was summed, as shown in the column Test 1 of Table 2. Here, a difference in dialect, for example, “uchira” (“we” in Osaka dialect) and “watashitachi” (“we” in Tokyo dialect), was permitted as correct recall. Each answer was relatively consistent. This result suggests that subjects tend to rearrange spoken contents following their own grammatical systems to memorize them, even if they are incomplete utterances. There was no subject who pointed out any unacceptability for the stimulus speech.

[Experiment-2]

The same procedure as for experiment-1 was performed using the stimulus (2) without pauses. Results are shown in column Test 2 of Table 2. The number of phrases recalled in the second experiment differs greatly from that in the first experiment. The difference is statistically significant at the 1% level. This indicates that pauses may affect the performance of the listener’s memory.
Table 2 The number of semantic elements recalled by groups 1 and 2.

<table>
<thead>
<tr>
<th>Phrase</th>
<th>Gloss</th>
<th>Test(1)</th>
<th>Test(2)</th>
</tr>
</thead>
<tbody>
<tr>
<td>1. uchira-no</td>
<td>our</td>
<td>7</td>
<td>4</td>
</tr>
<tr>
<td>2. kinjo-no</td>
<td>nearby</td>
<td>4</td>
<td>1</td>
</tr>
<tr>
<td>3. wakai-toki</td>
<td>young</td>
<td>4</td>
<td>4</td>
</tr>
<tr>
<td>4. densya-ga</td>
<td>streetcar</td>
<td>5</td>
<td>4</td>
</tr>
<tr>
<td>5. meiji</td>
<td>Meiji period</td>
<td>3</td>
<td>0</td>
</tr>
<tr>
<td>6. 37nenni</td>
<td>year 37</td>
<td>2</td>
<td>1</td>
</tr>
<tr>
<td>7. Yotsubashi</td>
<td>Yotsubashi</td>
<td>7</td>
<td>6</td>
</tr>
<tr>
<td>8. koosaten-de</td>
<td>intersection</td>
<td>6</td>
<td>2</td>
</tr>
<tr>
<td>9. mina</td>
<td>everyone</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>10. goza</td>
<td>grass mat</td>
<td>6</td>
<td>3</td>
</tr>
<tr>
<td>11. motte</td>
<td>bring</td>
<td>4</td>
<td>3</td>
</tr>
<tr>
<td>12. suwari-ni</td>
<td>sit</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>13. minuki-manneya</td>
<td>go to see</td>
<td>6</td>
<td>3</td>
</tr>
<tr>
<td>Total recalled by seven subjects each</td>
<td>54</td>
<td>32</td>
<td></td>
</tr>
</tbody>
</table>

4.3. Pauses and listener’s understanding

The result of the second experiment indicates that removing pauses may impair the capacity of the listener’s memory. To further study this problem, an additional experiment using speech materials from a TV news program (mentioned briefly in section 2.2) was planned. The same procedure as the second experiment was adopted, where all the pauses in the TV news were removed from the original speech material. The experimental result confirmed that listeners could not understand the messages lacking pauses (Sugito, 1989, 1994) [7].

A pause by itself does not convey any information. However, these results suggest that a pause is necessary for the speaker, for physiological reasons, for transmitting grammatical information on juncture, and for information retrieval in the brain for production. At the same time, for the listener, a pause is an indispensable period for the rehearsal in short-term memory and for reconstruction of the speaker’s message.

5. Conclusions

Transcribed text from conversational utterances is usually difficult to read and to understand, because it is often ill-formed and inconsistent from the viewpoint of grammar of written language. However, when we listen to the utterance as speech, we can understand it very well. The present study demonstrates that the prosodic characteristics of speech play very important roles for listener’s understanding.

(1) The experimental results on listener’s memorization of utterances indicate that the listener may understand speech by performing some content-reconstruction during the speaker’s pause intervals.

(2) A close relationship between pause length and preceding utterance length was not observed. At least in conversational speech, pauses are flexibly inserted at grammatical phrasal junctures, rather than obeying the physiological constraints of respiration.

(3) Utterance in hesitation or with truncation is brief, and the succeeding pause is also brief. The intonation in such speech has peculiar characteristics.

(4) Accented feature does not appear in an utterance while retrieving the next words in the brain, even if the word in the utterance is accented.

(5) Intonation in the low pitch range and with declination characteristics helps to elicit a brief response or a back-channel comment from the listener. In addition to these, slowing down of speech tempo introduces a chance for turn-taking in the discourse.

(6) Intonation of a phrase without an important meaning has a flat pattern and remains in the low pitch range. Important passages are distinguishable for the listeners, because their pitch frequency patterns of accent and intonation are clearly different from those of the less meaningful passages.

In conversational speech, pause, declining intonation and tempo slowdown are signs of breaks in discourse. In important or meaningful phrases, the accentual and intonational characteristics are distinctly produced. Thus, it is considered that ill-formed utterances can be memorized according to the listener’s grammatical system, and are easily understood. The flat intonation characteristics are observed not only in hesitation, but also during an utterance while the next word is being retrieved. It is supposed that listeners utilize such differences in intonation patterns to sort important information.

Taken together, these findings suggest that utterances of spoken language are not really ill-formed, but constitute legitimate, well-formed messages which the listener can understand without difficulty. The use of prosodic features both by the speaker and by the listener is an important part of the grammar of spoken language, which is quite distinct from that of written language. It is hoped that the present study will serve as a preliminary step toward finding and constructing a grammar of spoken language.

References


