A Technique for Determining Time Length of Pause at Final Period of Each Sentence in Story: Design of Synthesized Speech for Supporting Children’s Comprehension

Hideji Enokidu† Kazuhiro Uenosono‡ Seiichi Komiya‡
†Shibaura Institute of Technology
3-7-5 Toyosu, Koutou-ku, Tokyo, 135-8548 Japan
‡Graduate School of Engineering, Shibaura Institute of Technology
3-7-5 Toyosu, Koutou-ku, Tokyo, 135-8548 Japan
†enokizu@shibaura-it.ac.jp, ‡{m704101, skomiya}@shibaura-it.ac.jp

Abstract: The present study tried to clarify that the storyteller controlled the time length of pause at the final period in each sentence to support children’s comprehension and suggest the way in which given findings are applied to reading the text by the use of synthesized speech. During listening to the sound of telling a story text, a listener constructs a mental microworld that is refered to as the term situation model in the discourse psychology. Unlike the storyteller who can read the text at his own pace, however, the listener’s construction of the situation model should be influenced by the way in which a storyteller reads the text. Because the event-indexing model indicated that comprehending time of a sentence in a story is progressively longer to the extent that there are continuity breaks on more dimensions (e.g., space, time, characters and objects), we focused on the relationship between the number of mental operations to comprehend a sentence and the time length of pause at the final period in each sentence of the story read by the storyteller. In order to estimate the mental operations that are required to construct the situation model, we devised the multilayered frame representation. Then, we defined a sequence of the operations to transform the case frame representation of each sentence to the multilayered frame representation and examined the relationship between it and the time length of pause at period in the speech sound of a vastly-experienced storyteller. The results indicated that the storyteller controlled the time length of pause according to the number of the operations so as to help the listener to comprehend the story.

Key-word: discourse comprehension, situation model, event-indexing model, multilayered frame representation, storyteller, time length of pause at final period, application to speech synthesis

1 Introduction

Needless to say, the prosodical cues of speech sound in telling the story are relevant to the listener’s comprehension. We obtained several findings that imply the relations between the development of stories and the time length of the pause, the character’s emotion and the pitch, and the story structure and the speed. Researches, however, tend to examine them without discussing fundamental problem such as what comprehension is and how it is performed in the listener’s mind [1][2]. Among the prosodical cues, the pause at the final period is suggested to have close relationship to the comprehension of each sentence particularly.

The silent pause at final period (that is the pause at the sentence boundary), which is referred by the present study, is especially focused on in the researches of the listener’s comprehension, because it is closely related to the semantic processing of each sentence and the story structure. The research found that the radio announcer usually sets about 1400msec pause at final period in each sentence when he or she reads the news article and the weather report [3]. However, such pause length is not necessarily most suitable for listener to comprehend them. Other researches indicated that experienced readers set the pause of varied time lengths in reading story texts [4][5]. Thus, these evidences clearly indicate that the constant time length at final period is not appropriate to support the listener’s comprehension. In fact, it was suggested that the appropriate pause length, which can facilitate the listener’s comprehension, is varied according to the story or the text structure [6]. It was also demonstrated that storyteller increased the pause length at the sentence boundary which occupies critical place in the text structure [7]. Although these researches examined only the relation between the surface structure of the text and the pause length, they didn’t discuss the model of the listener’s comprehension substantially.
Some cognitive linguists conceived of language as a set of cues by which the speaker manipulates the listener’s attention on the fictional situation [8]. According to this suggestion, the storyteller reads the story for the listener to construct the microworld described by the sentence and to be immarced in it as shown in Fig. 1. In the present study, we proposed the mental representation of the microworld, which is comparable to the knowledge representation in AI, in the listener’s mind. Then it was assumed that the storyteller controls the time length of the pause at the final period in each sentence of the story to support the listener’s comprehension. Especially, if the listener is a young child, it is true. Therefore, if the pause length at final period set by the storyteller reflects to the construction of our mental representation, it is likely to indicate some aspects of the mental situation that is actually constructed in the listener’s mind.

The purpose of our study is primarily to examine whether the construction processes of our mental presentation actually simulate the mental processes on comprehending the story. And furthermore, we tried to suggest how to determine the time length of the pause at final period of each sentence in the text that is read by the synthesized speech.

2 Discourse Comprehension

Discourse comprehension is thought to be a series of transforming the discourse representation. Most discourse psychologists especially adopt van Dijk and Kintsch’s distinction[9] among the surface code, the textbase, and the situation model. The surface code preserves the exact wording and syntax of clauses. In a way, it may imply the concept of each word and the relation among such concepts. The textbase contain the explicit text propositions in a refined form that preserve the meaning of each clause. Also the textbase includes a little number of inferences needed to maintain the local text coherence. The situation model is the microworld described by each clause and in whole text. The situation model refers to the setting, the people, the object, and action involved in each event that occurs in the mental microworld.

Historically prevalent theory of the mental representation such as the textbase and the situation model has premised the amodal symbol system that is conceptual and propositional [10][11][12][13]. However the perspective on the discourse comprehension is influenced by the development of the cognitive linguistics and the brain science. In particular, theoretical implications and empirical evidences obtained by neuroimaging (e.g., fMRI) researches in the brain science dramatically impact the conventional perspective on mental operations underlying the cognition process common to various human activities. Based on the findings given by the brain science, Barsalou has recently presented alternative to the amodal symbolic system in the form of Perceptual Symbol System[14][15]. Barsalou’s theory assumes that modal symbolic system in the brain supports diverse forms of the mental simulation across different cognitive processes (e.g., perceptual, motor, and affective simulations). Furthermore, Barsalou integrated his theory with the situated cognition, proposing that simulations typically embody the concepts involved in the situation model such as characters, objects, actions, events, and mental state [16]. Perceptual Symbol System obviously impacted on the research of the situational model in discourse comprehension.

2.1 Situation Model

Until the early 1980s, many cognitive psychologists viewed the discourse comprehension as the construction and retrieval of the mental representation of the text itself in the discourse genre. This perspective was changed by two seminal books published in 1983 [17][9]. Their ideas about the mental representation of situation described verbally, which has become to be known as the mental model [17] or the situation model [9], has still underlay on the amodal symbol system such as concepts and propositions. However, the great shift has recently occurred in thinking about the role of language and the modality of the situation model.

Rather than treating language as information to
analyze syntactically and semantically and then store as the concept and the proposition in memory. This perspective might reflect the salient rise of the cognitive linguistics. In addition to new perspective of the language role, the empirical evidence of the brain science and several suggestions given by Perceptual Symbol System made the notion of the discourse comprehender as an immersed experimenter in the mental microworld [19][20][21]. Because this view suggests both the speaker and the listener actually experience information in a discourse as if they were participating in the activity, it is consistent with the notion of the comprehension as mental simulation. Additionally it is also consistent with other views of embodied cognition, according to which cognition is grounded in perception and action and relies on the use of perceptual and motor representations rather than of abstract, amodal, and arbitrary mental representations such as propositional networks or feature lists.

The empirical evidences have amassed recently showing that visual representations are usually activated during discourse comprehension. This includes visual representation of the object shape [22][23], orientation [24], motor direction [25]. These evidences suggest that the situation model is the mental representation which permits the reader or the listener to simulate the entities and the event described by the sentence perceptually.

Additionally many researchers obtained the empirical evidence that motor representations are also activated during discourse comprehension. In the area of the brain science, Pulvermüller found that when participants read the word for an action, the motor system becomes active to represent its meaning [26]. Other researchers have assessed whether physical actions affect the sentence comprehension using behavioral measures. Klatsky et al. showed the priming effect of a motor action on the time to judge the sensibility of a simple phrase describing an action [27]. Similarly, comprehension is facilitated when the action to make a response is consistent with text meaning [28] and also when the action to control text presentation is consistent [29]. These findings implied that the situation model is the mental representation which permits the reader or the listener to do motor simulation of the action described by the sentence.

In consideration of those findings, it should be specifically examined how to represent the situation model as the mental representation. The way to represent the situation model, which will be proposed later, reflected the empirical evidence given by the recent studies.

2.2 Processes for Constructing Situation Model

To the present, several researchers have proposed the theory or the model that explains how to construct the situation model. For example, the structure-building framework [18], the construction-integration model [30], the landscape model [31], and the event-indexing model [32][33][34] are representative ones. On the basis of comparison among them, the event-indexing model was selected as theoretical foundation. The determinative reasons for selection come from the fact that it specified the cognitive processes that are likely to predict the time for reading each sentence involved in the text.

The event-indexing model makes claims about both on-line comprehension and the mental representation resulting on the reader’s long-term memory. According the model, on reading the story, each event described in the sentence is decomposed into five indexes: time, space, causality, intentionality (character’s goal), and character/object. These dimensions correspond to the dimensions listed by Chafe [35]. Then the coherent situation model is constructed by connection with shared indexes between the currently processed event and the constructed situation model until now.

Incoming events can be more easily connected with evolving situation model to the extent that they share indexes with the current state of the situation model. Thus, the event-indexing model makes the general prediction that the processing load during comprehension varies as a function of the number of situational indexes shared between the currently processed event and the current state of the situation model. This processing load hypothesis was supported by several studies [32][36][37].

3 Multilayered Frame Representation of Situation Model

Zwaan and Radvansky also tried to use the frame concept to represent the situation model as one type of knowledge representation[34]. Since the frame is originated to represent the knowledge of visual world by Minsky [38], it seems to be appropriate to represent the situation model. They assumed that the establishment of the spatio-temporal frame is obligatory during the construction of the situation model so that it
grounds situations in space and time. And furthermore, they referred to the entity frames (the character and the object frames) associated with physical and mental attributes. Although they didn’t indicate so with certainty, it is appeared that they regarded the situation model as the knowledge representation with the hierarchical structure setting spatio-temporal frame set at a top phase.

According their suggestion, we regarded the situation model as the multilayered frame representation containing the constituents of the event described in the story. Fig. 2 shows the components and structure of the multilayered frame representation named as the situation frame. The situation frame will be explained in a little detail below.

![Configuration of multilayered situation frame.](image)

The reader extracts the meaning from each sentence of the story using lexical and grammatical knowledge. Such meaning was assumed to be composed of the predicate, the character, who involved the personalized animal, object, the instrument, the source, the goal, the time concepts and so on according to the case frame representation [39]. In the present study, the components of the meaning were inputs for constructing the situation frame. It is explained below how to generate several kinds of frame involved in the situation frame and to insert the value into slots equipped each frame with.

**Scene Frame** When the predicate concept is the verb relevant to transfer of the character from one space to another space and the goal and/or the time concepts refer to specific values, the scene frame is generated. The scene frame has the space, the time, the character, and the objects slots. On generating the scene frame, the goal and/or the time concepts are inserted into their correspondent slot. And when the character and/or the object concepts have specific values as inputs at first time, they are inserted into their correspondent slots. The values in the character and the object slots are also the reference point of their correspondent frames.

**Character Frame** When a value of the character concept is first inserted into the character slot in the scene frame, the character frame is generated. At the same time, the value is inserted into the name slot. Then, the value of the predicate concept referred to the physical or psychological attribute of the character is inserted into the correspondent slot after the value of the character concept is matched against the value that has been inserted into the character slot of the scene frame already. At the same time when the value of the predicate concept referred to the action accompanied with the physical movement or the utterance is inserted into the action slot, the frame according to it is generated.

**Object Frame** When a value of the object concept is first into the object slot, the object frame is generated. And the value is inserted in to the name slot of the object frame. Once the object frame is generated, the value of the object concept is matched against the value in the object slot and inserted into the name slot.

An action accompanied with the physical attribute of the object is inserted into the slot of the physical attribute. On the other hand, when the predicate concept is relevant to the action, its value is inserted in to the action slot and instigates to generate the action frame. If the action frame is identical with the action frame generated from the character frame at the same moment, their frame are unified.

**Action and Utterance Frames** As explained in character and object frames, the action frame is generated by the predicate concept referred with the action accompanied with physical movement, and the utterance frame is generated by the predicate referred to the action relative to utterance (e.g., say, claim, murmur, and so on). On the utterance frame, contents of the utterance are written simultaneously.

Remember that we intend to devise the modal representation of the events described in the story, told in the conversation, or directly observed in daily life. Therefore, the scene, the character, and the object frame are the perceptual representation. On the other hand, the action and the utterance frame is the motoric representation.

## 4 Method

### 4.1 Stories for Reading and Storyteller

In this study, we used five stories that were involved in two illustrated books published for reading to the children. Those five stories were all classical Japanese fairly stories as follows: “The Stone Buddhas”, “A Crane’s Gratitude”, “The Rolling Rice Ball”, “The Princess of the Moon”, and “Click Click Mountain”[40][41].
Ms. H. T. was asked to be a storyteller for reading five fairy stories to several undergraduate students. For nearly two decades, she participates in the volunteer activity that tries to contribute to the early childhood education by reading the folk story for children.

4.2 Procedure
Measuring the Time Length of Pause at Final Period of Each Sentence
The storyteller’s speech sound of reading each story was recorded with the digital recorder (TOSHIBA DMR-SX2). The speech sound was stored onto the USB memory in WAVE file format. Using the SUGI speech analyzer (ANIMO ANMSW-SSA0101), two cooperators measured the time length of the pause at final period of each sentence in millisecond. Their measured values were averaged to be input data for the explanatory variable in the multiple linear regression analysis.

4.3 Coding of Mental Operations for Constructing the Situation Frame
It is assumed that construction of the situation frame is equivalent to transformation from the sentence meaning to the situation model. As mentioned before, on constructing the situation frame, the several kinds of the mental operation are needed to the sentence meaning. In order to determine what kinds of the mental operation are carried out at the time, we simulated the processes of constructing the situation frame. At first, we used the propositional notation that was suggested by Fillmore [39] and developed by Kintsch [42] to represent the meaning of each sentence. Then, we transformed such propositional notation to the situation frame manually to estimate the kinds of mental operation during comprehension of each sentence. In this way, we gave the code to the mental operations that were carried out to construct the situation frame.

Table 1 presented the codes to be referred to each kind of the mental operation. For example, Table 2 indicates the codes given to the some sentences and the pause length measured at their final period.

5 Results
We set the pause length as the objective variable and the mental operations as the explanatory variables to conduct the multiple regression analysis. The total number of the mental operations carried out to construct the situation frame is sixteen as shown in Table 3.

<table>
<thead>
<tr>
<th>Codes Given to Mental Operations to Construct Situation Frame</th>
</tr>
</thead>
<tbody>
<tr>
<td>GS0: Generation of Scene Frame</td>
</tr>
<tr>
<td>WTS: Writing Time Concept in Time Slot</td>
</tr>
<tr>
<td>WSS: Writing Space Concept in Space Slot</td>
</tr>
<tr>
<td>WGSCn: Writing Character Concept in Character Slot and Generation of Character Frame</td>
</tr>
<tr>
<td>WGSO:n: Writing Object Concept in Character Slot and Generating Object Frame</td>
</tr>
<tr>
<td>WGCA: Writing Action Concept in Action Slot and Generating Action Frame</td>
</tr>
<tr>
<td>WGCU: Generation of Utterance Frame and Writing Utterance Content</td>
</tr>
<tr>
<td>WGOn: Writing Physical Attribute Concept in Physical Attribute Slot</td>
</tr>
<tr>
<td>WGOA: Writing Action Concept in Action Slot and Generation Action Frame</td>
</tr>
<tr>
<td>Other Codes</td>
</tr>
<tr>
<td>GSH: Preparation of Generating Scene Frame</td>
</tr>
<tr>
<td>UF: Binding of Character Frame or Object Frame</td>
</tr>
</tbody>
</table>

Note m:0=previously given frame, 1=generated frame during comprehension
n: 0=original frame, 1=recalled frame, 2=changing or adding frame name

On the basis of the number of the mental operations when the case-frame representation is transformed to the multilayered situation frame, the value of the explanatory variables is determined. Therefore, the value of the variables corresponding to the generation of each frame is 0 or 1, and the value of the variables cor-
responding to the insertion into each slot is from one to four.

When the full regression method was used, the results of the multiple regression analysis indicated that sixteen explanatory variable can predict the pause length at the final period of each sentence with considerable accuracy \((F(16,50)=4.660, p<.001;\) determination coef., .599; multiple correlation coef., .774; Durbin-Watson ratio, 2.172). Specifically, GSH, WGSO0, and WGCU were significant explanatory variables. Fig. 3 presents the relationship between the theoretical value and the observation value. These results, however, might show the inflated predictable power of the multiple regression equation because of the number of the explanatory variables.

Consequently, the multiple regression analysis was conducted again by the use of the stepwise regression method. The results indicated that GS1, GSH, WGSO0, WGSO2, and WGCU influenced on the pause length significantly as shown in Table 4. The result of ANOVA \((F(5,61)=15.352, p<.001)\) and the measures of prediction accuracy indicated that only five explanatory variables were sufficient to predict the pause length. Table 5 shows several measures of prediction accuracy specifically. Fig. 4 shows the relationship between the relationship between the theoretical value and the observation value. These results indicate that the multiple regression equation with five variables and constant can predict the final pause considerably.

**Table 4**

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>WGSO0</td>
<td>223.72</td>
<td>16.765</td>
<td>4.094</td>
<td>0.001**</td>
<td>0.454</td>
<td>0.509</td>
</tr>
<tr>
<td>GSH</td>
<td>487.89</td>
<td>34.544</td>
<td>5.877</td>
<td>0.001**</td>
<td>0.601</td>
<td>0.372</td>
</tr>
<tr>
<td>WGC</td>
<td>318.41</td>
<td>20.570</td>
<td>4.535</td>
<td>0.001**</td>
<td>0.502</td>
<td>0.178</td>
</tr>
<tr>
<td>GS1</td>
<td>310.55</td>
<td>12.782</td>
<td>3.575</td>
<td>0.001**</td>
<td>0.416</td>
<td>0.066</td>
</tr>
<tr>
<td>WGSO2</td>
<td>414.38</td>
<td>3.904</td>
<td>1.976</td>
<td>0.053+</td>
<td>0.245</td>
<td>0.196</td>
</tr>
<tr>
<td>Constant</td>
<td>1230.85</td>
<td>363.757</td>
<td>19.072</td>
<td>0.001**</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

**Table 5**

<table>
<thead>
<tr>
<th>Several Measures of Prediction Accuracy</th>
</tr>
</thead>
<tbody>
<tr>
<td>Determination coef.</td>
</tr>
<tr>
<td>Adjusted determination coef.</td>
</tr>
<tr>
<td>Multiple correlation coef.</td>
</tr>
<tr>
<td>Adjusted multiple correlation coef.</td>
</tr>
<tr>
<td>Durbin-Watson ratio</td>
</tr>
</tbody>
</table>
6 Discussion

The situation frame is a devised knowledge representation so as to specify the situation model. Therefore, if the storyteller wants the listeners to comprehend the story, she must facilitate their construction of the situation model by the use of diverse ways. As described earlier, one of such ways was hypothesized to be the control of the pause length in final period of each sentence. The results of the multiple regression analysis reasonably supported this hypothesis. Particularly, it is important that the significant explanatory variables, which have an influence on the pause length, are related to the frame generation. According to the processing load hypothesis, when the reader comprehends the story, reading time of each sentence increases in proportion to the number of the dimensional discontinuities between the integrated and the current situation models. However, because it is not the reader but the listener that comprehends the story in present study, the storyteller must read it in consideration of such processing load. That is, it is possible to say that my experienced storyteller intentionally or unintentionally controlled the pause length at the final period to help the listener construct the situation model. Then we will examine more specifically the results and finally indicate the implications to the interpersonal communication.

6.1 Effect of Scene Frame Generation

The results indicated that the generation of the scene frame at the set of new scene increases the pause length by 311msec except for the generation of the scene frame at the first time when the situation frame is constructed. In addition, it was indicated that the generation the scene frame based on the prospective description increases the pause length by 488msec.

These results are critically important that the scene frame is assumed to embody the concepts that compose the event described by each sentence. Therefore, it is implied that she increased the pause length to help the listener embody the object concept and simulate the event relevant to it. Changing the name of the object frame also increases it by 224msec. And change of the object frame name also increases it by 414msec.

At first glance, these results are appeared to be strange because each character play an important role in the story. The character frames may be ready to generate so that the characters in used stories as the materials are prototypical ones in the folk story (e.g., grandpa, grandma, young hero, princess, and so on). So it is assumed that the characters in the story are embodied by the listener on hearing the storyteller’s reading.

On the other hand, it is thought that each object frame is not easily generated so that the objects involved in the folk story are so diverse and sometimes unusual. If the storyteller was permitted to communicate wit gestures, she probably used her hands, or show and draw the picture to embody the described object. When it is unusual, she was sure to do so. Therefore, it is implied that she increased the pause length to help the listener embody the object concept and simulate the event relevant to it. The estimation of such processing load might induce to greater increase of the pause length.

6.2 Effects of Character and Object Frame Generation

Any generation of the character frame didn’t have an effect on the pause length at all. However, generation of the object frame and change (or addition) of the object frame name have significant and marginally significant effect respectively. Generation of the object frame increases the pause length by 224msec. And change of the object frame name also increases it by 414msec.

6.3 Effect of Utterance Frame Generation

The utterance frame in the third layer of the situation frame is originally set on the basis of difference of the sentence type in the story. The generation of the utterance frame increased the pause length by 318msec. Unlike the narrative sentence, the conversational
The results, which were given by the multiple regression analysis, reflect the increase of the pause length. It may also suggest that the increase of the pause length on generating the utterance frame is reflected on the time that the storyteller needs to tune her voice to the character’s voice quality.

6.4 Effects of Action Frame Generation and Value Insertion
Both kinds of multiple regression analysis didn’t indicate the significant effect of the generation of the action frame. The action frame is set in the third layer and has particularly strong tie with the character frame that has little effects on the pause length. Therefore, the action frame may be generated together with the character frame on hearing the reading of each sentence. The generation of the action frame must be closely examined in the future study because the two critical indices (i.e., intention and causality) of the described event, which were not treated in the present study, are included in the action.

The insertion of value in any slots had little effect on the increase of the pause length. Since the insertion means the association with the frame, it may not take time as to need the increase of the pause length.

6.5 Contribution to Speech Synthesis System
The results, which were given by the multiple regression analysis using the stepwise regression method, suggested that the time length at the final period of each sentence can be predicted by some kinds of the frame generation on constructing the multilayered frame representation. Although the storyteller participating in this study is only one, the time length at the final period can be set by the use of the equation obtained in our study as if the experienced storyteller designs it.

7 Conclusion
We generally communicate our daily experiences each other. Most of them are likely to be the events in which we involved as the participants and the eyewitnesses. From such a viewpoint, sending a message in the communication takes the form of the discourse. Then in the communication situation, the storyteller must try to carry his or her own experiences to the listener in one way or another. That is, the storyteller reads the message with consideration for the listener’s comprehension of it. The pause length set at the final period of each sentence is likely to be one of the considerations because it is thought to help the listener extract meaning from the sentence.

In the area of the psycholinguistics, it is assumed that the listener constructs the situation model to comprehend the discourse. The situation model is the mental representation of the event occurring in the microworld described in the discourse. Recently, the researches in the area of brain science indicated that the situation model is a modal representation permitting the listener to simulate it mentally. In the present study, we proposed the knowledge representation named as the situation frame to specify the situation model. The situation frame has multilayered frame structure composed of the scene, the character, the object, the action, and the utterance frames.

The present study investigates the intrapersonal processes through which the storyteller sends a message to the listener permitting to construct the situation frame. Specifically, we asked the storyteller to read the folk tales as if she was facing the younger children. Then, the multiple regression analysis was conducted to the pause length set by the storyteller with the multiple regression analysis to examine whether it reflects on the mental operations needed for constructing the situation frame. The results generally indicated that the increase of the pause length can be predicted by such mental operations.

The situation frame, however, is still not sufficient to represent the mental representation for simulating the event. Currently we are trying to refine the situation frame so that the described event can be transformed into animation.

References:


R. A. Zwaan, & D. N. Rapp, Discourse comprehension, In M. J. Traxler & M. A. Gernsbacher (Eds.), Handbook of Psycholinguistics, 2nd edition (pp.725-764), New York: Academic Press, 2006


P. Van den Broek, K. Young, Y. Tzeng, & T.


