Converting Text into Agent Animations: Assigning Gestures to Text

Yukiko I. Nakano[†] **Masashi Okamoto**[‡] **Daisuke Kawahara**[‡] **Qing Li**[‡] **Toyoaki Nishida**[‡]

[†]Japan Science and Technology Agency
2-5-1 Atago, Minato-ku, Tokyo, 105-6218 Japan

[‡]The University of Tokyo
7-3-1 Hongo, Bunkyo-ku, Tokyo, 113-8656 Japan

{nakano, okamoto, kawahara, liqing, nishida}@kc.t.u-tokyo.ac.jp

Abstract

This paper proposes a method for assigning gestures to text based on lexical and syntactic information. First, our empirical study identified lexical and syntactic information strongly correlated with gesture occurrence and suggested that syntactic structure is more useful for judging gesture occurrence than local syntactic cues. Based on the empirical results, we have implemented a system that converts text into an animated agent that gestures and speaks synchronously.

1 Introduction

The significant advances in computer graphics over the last decade have improved the expressiveness of animated characters and have promoted research on interface agents, which serve as mediators of human-computer interactions. As an interface agent has an embodied figure, it can use its face and body to display nonverbal behaviors while speaking.

Previous studies in human communication suggest that gestures in particular contribute to better understanding of speech. About 90% of all gestures by speakers occur when the speaker is actually uttering something (McNeill, 1992). Experimental studies have shown that spoken sentences are heard twice as accurately when they are presented along with a gesture (Berger & Popelka, 1971). Comprehension of a description accompanied by gestures is better than that accompanied by only the speaker's face and lip movements (Rogers, 1978). These previous studies suggest that generating appropriate gestures synchronized with speech is a promising approach to improving the performance of interface agents. In previous studies of multimodal generation, gestures were determined according to the instruction content (Andre, Rist, & Muller, 1999; Rickel & Johnson, 1999), the task situation in a learning environment (Lester, Stone, & Stelling, 1999), or the agent's communicative goal in conversation (Cassell et al., 1994; Cassell, Stone, & Yan, 2000). These approaches, however, require the contents developer (e.g., a school teacher designing teaching materials) to be skilled at describing semantic and pragmatic relations in logical form. A different approach, (Cassell, Vilhjalmsson, & Bickmore, 2001) proposes a toolkit that takes plain text as input and automatically suggests a sequence of agent behaviors synchronized with the synthesized speech. However, there has been little work in computational linguistics on how to identify and extract linguistic information in text in order to generate gestures.

Our study has addressed these issues by considering two questions. (1) Is the lexical and syntactic information in text useful for generating meaningful gestures? (2) If so, how can the information be extracted from the text and exploited in a gesture decision mechanism in an interface agent? Our goal is to develop a media conversion technique that generates agent animations synchronized with speech from plain text.

This paper is organized as follows. The next section reviews theoretical issues about the relationships between gestures and syntactic information. The empirical study we conducted based on these issues is described in Sec. 3. In Sec. 4 we describe the implementation of our presentation agent system, and in the last section we discuss future directions.

2 Linguistic Theories and Gesture Studies

In this section we review linguistic theories and discuss the relationship between gesture occurrence and syntactic information.

Linguistic quantity for reference: McNeill (McNeill, 1992) used communicative dynamism (CD), which represents the extent to which the message at a given point is 'pushing the communication forward' (Firbas, 1971), as a variable that correlates with gesture occurrence. The greater the CD, the more probable the occurrence of a gesture. As a measure of CD, McNeill chose the amount of linguistic material used to make the reference (Givon, 1985). Pronouns have less CD than full nominal phrases (NPs), which have less CD than modified full NPs. This implies that the CD can be estimated by looking at the syntactic structure of a sentence.

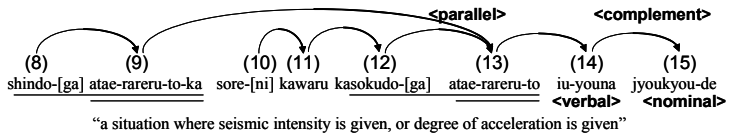

Figure 1: Example analysis of syntactic dependency
Underlined phrases are accompanied by gestures, and strokes occur at double-underlined parts. Case markers are enclosed by square brackets [ ].

Theme/Rheme: McNeill also asserted that the theme (Halliday, 1967) of a sentence usually has the least CD and is not normally accompanied by a gesture. Gestures usually accompany the rhemes, which are the elements of a sentence that plausibly contribute information about the theme, and thus have greater CD. In Japanese grammar there is a device for marking the theme explicitly. Topic marking postpositions (or "topic markers"), typically "wa," mark a nominal phrase as the theme. This facilitates the use of syntactic analysis to identify the theme of a sentence. Another interesting aspect of information structure is that in English grammar, a wh-interrogative (what, how, etc.) at the beginning of a sentence marks the theme and indicates that the content of the theme is the focus (Halliday, 1967). However, we do not know whether such a special type of theme is more likely to co-occur with a gesture or not.

Given/New: Given and new information demonstrate an aspect of theme and rheme. Given information usually has a low degree of rhematicity, while new information has a high degree. This implies that rhematicity can be estimated by determining whether the NP is the first mention (i.e., new information) or has already been mentioned (i.e., old or given information).

Contrastive relationship: Prevost (1996) reported that intonational accent is often used to mark an explicit contrast among the salient discourse entities. On the basis of this finding and Kendon's theory about the relationship between intonation phrases and gesture placements (Kendon, 1972), Cassell & Prevost (1996) developed a method for generating contrastive gestures from a semantic representation. In syntactic analysis, a contrastive relation is usually expressed as a coordination, which is a syntactic structure including at least two conjuncts linked by a conjunction.

Figure 1 shows an example of the correlation between gesture occurrence and the dependency structure of a Japanese sentence. Bunsetsu units (8)-(9) and (10)-(13) in the figure are conjuncts. A "bunsetsu unit" in Japanese corresponds to a phrase in English, such as a noun phrase or a prepositional phrase. Each conjunct is accompanied by a gesture. Bunsetsu (14) is a complement containing a verbal phrase; it depends on bunsetsu (15), which is an NP. Thus, bunsetsu (15) is a modified full NP and thus has large linguistic quantity.

3 Empirical Study

To identify linguistic features that might be useful for judging gesture occurrence, we videotaped seven presentation talks and transcribed three minutes for each of them. The collected data included 2124 bunsetsu units and 343 gestures.

Gesture Annotation: Three coders discussed how to code the half the data and reached a consensus on gesture occurrence. After this consensus on the coding scheme was established[1], one of the coders annotated the rest of the data. A gesture consists of preparation, stroke, and retraction (McNeill, 1992), and a stroke co-occurs with the most prominent syllable (Kendon, 1972). Thus, we annotated the stroke time as well as the start and end time of each gesture.

Linguistic Analysis: Each bunsetsu unit was automatically annotated with linguistic information using a Japanese syntactic analyzer (Kurohashi & Nagao, 1994)[2]. The information was determined by asked the following questions for each bunsetsu unit.

(a) If it is an NP, is it modified by a clause or a complement?
(b) If it is an NP, what type of postpositional particle marks its end (e.g., "wa", "ga", "wo")?
(c) Is it a wh-interrogative?
(d) Are all the content words in the bunsetsu unit have mentioned in a preceding sentence?
(e) Is it a constituent of a coordination?

Moreover, as we noticed that some lexical entities frequently co-occurred with a gesture in our data, we used the syntactic analyzer to annotate additional lexical information based on the following questions.

(f) Is the bunsetsu unit an emphatic adverbial phrase (e.g., very, extremely), or is it modified by a preceding emphatic adverb (e.g., very <u>important</u> issue)?
(g) Does it include a cue word (e.g., now, therefore)?
(h) Does it include a numeral (e.g., <u>thousands</u> of people, <u>99</u> times)?

We then investigated the correlation between these lexical and syntactic features and the occurrence of gesture strokes.

Result: The results are summarized in Table 1. The baseline gesture occurrence frequency was 10.1% per bunsetsu unit (a gesture occurred once about every ten

[1] Inter-coder reliability among the three coders in categorizing the gestures (beat, iconic, etc.) was sufficiently high (Kappa = 0.81). Although we did not measure agreement on gesture occurrence itself, this result suggests that the coders had very similar schemes for recognizing gestures.

[2] To prevent the effects of parsing errors, errors in syntactic dependency analysis were corrected manually for about 13% of the data.

Table 1. Summary of results

Case	Syntactic/lexical information of a bunsetsu unit			Gesture occurrence
C1	Quantity of modification	(a) NP modified by clause		0.382
C2		Pronouns, other types of NPs	(b) Case marker = "wo" & (d) New information	0.281
C3	(c) WH-interrogative			0.414
C4	(e) Coordination			0.477
C5	Emphatic adverbial phrase	(f) Emphatic adverb itself		0.244
C6		(f') Following emphatic adverb		0.350
C7	(g) Cue word			0.415
C8	(h) Numeral			0.393
C9	Other (baseline)			0.101

bunsetsu units). A gesture stroke most frequently co-occurred with a bunsetsu unit forming a coordination (47.7%). When an NP was modified by a full clause, it was accompanied by a gesture 38.2% of the time. For the other types of noun phrases, including pronouns, when an accusative case marked with case marker "wo" was new information (i.e., it was not mentioned in a previous sentence), a gesture co-occurred with the phrase 28.1% of the time. Moreover, gesture strokes frequently co-occurred with wh-interrogatives (41.4%), cue words (41.5%), and numeral words (39.3%). Gesture strokes frequently occurred right after emphatic adverbs (35%) rather than with the adverb (24.4%).

These cases listed in Table 1 had a 3 to 5 times higher probability of gesture occurrence than the baseline and accounted for 75% of all the gestures observed in the data. Our results suggest that these types of lexical and syntactic information can be used to distinguish between where a gesture should be assigned and where one should not be assigned. They also indicate that the syntactic structure of a sentence more strongly affects gesture occurrence than theme or rheme and than given or new information specified by local grammatical cues, such as topic markers and case markers.

4 System Implementation

4.1 Overview

We used our results to build a presentation agent system, SPOC (Stream-oriented Public Opinion Channel)." This system enables a user to embody a story (written text) as a multimodal presentation featuring video, graphics, speech, and character animation. A snapshot of the SPOC viewer is shown in Figure 2.

In order to implement a storyteller in SPOC, we developed an agent behavior generation system we call "CAST (Conversational Agent System for neTwork applications)." Taking text input, CAST automatically selects agent gestures and other nonverbal behaviors, calculates an animation schedule, and produces synthesized voice output for the agent. As shown in Figure 2,

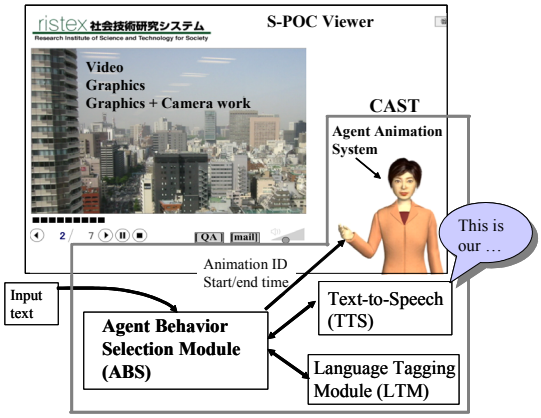

Figure 2: Overview of CAST and SPOC

CAST consists of four main components: (1) the Agent Behavior Selection Module (ABS), (2) the Language Tagging Module (LTM), (3) the agent animation system, and (4) a text-to-speech engine (TTS). The received text input is first sent to the ABS. The ABS selects appropriate gestures and facial expressions based on the linguistic information calculated by the LTM. It then obtains the timing information from the TTS and calculates a time schedule for the set of agent actions. The output from the ABS is a set of animation instructions that can be interpreted and executed by the agent animation system.

4.2 Determining Agent Behaviors

Tagging linguistic information: First, the LTM parses the input text and calculates the linguistic information described in Sec. 3. For example, bunsetsu (9) in Figure 1 has the following feature set.

{Text-ID: 1, Sentence-ID: 1, Bunsetsu-ID: 9, Govern: 8, Depend-on: 13, Phrase-type: VP, Linguistic-quantity: NA, Case-marker: NA, WH-interrogative: false, Given/New: new, Coordinate-with: 13, Emphatic-Adv: false, Cue-Word: false, Numeral: false}

The text ID of this bunsetsu unit is 1, the sentence ID is 1, the bunsetsu ID is 9. This bunsetsu governs bunsetsu 8 and depends on bunsetsu 13. It conveys new information and, together with bunsetsu 13, forms a parallel phrase.

Assigning gestures: Then, for each bunsetsu unit, the ABS decides whether to assign a gesture or not based on the empirical results shown in Table 1. For example, bunsetsu unit (9) shown above matches case C4 in Table 1, where a bunsetsu unit is a constituent of coordination. In this case, the system assigns a gesture to the bunsetsu with 47.7 % probability. In the current implementation, if a specific gesture for an emphasized concept is defined in the gesture animation library (e.g., a gesture animation expressing "big"), it is preferred to a "beat gesture" (a simple flick of the hand or fingers up and down (McNeill, 1992)). If a specific gesture is not defined, a beat gesture is used as the default.

The output of the ABS is stored in XML format. The type of action and the start and end times of the action are indicated by XML tags. In the example shown in Figure 3, the agent first gazes towards the user. It then performs contrast gestures at the second and sixth bunsetsu units and a beat gesture at the eighth bunsetsu unit.

Finally, the ABS transforms the XML into a time schedule by accessing the TTS engine and estimating the phoneme and bunsetsu boundary timings. The scheduling technique is similar to that described by (Cassell et al., 2001). The ABS also assigns visemes for the lip-sync and the facial expressions, such as head movement, eye gaze, blink, and eyebrow movement.

```
<Gaze type="towards">
    shindo-ga
        <Gesture_right type="contrast" handshape_right="stroke1@2">
        atae-rareru-to-ka
        </Gesture_right>
    sore-ni
    kawaru
    kasokudo-ga
        <Gesture_right type="contrast" handshape_right="stroke2@2">
        atae-rareru-to
        </Gesture_right>
    iu-youna
        <Gesture_right type="best" handshape_right="stroke1">
        jyoukyou-de
        </Gesture_right>
    ...
```

Figure 3: Example of CAST output

5 Discussion and Conclusion

We have addressed the issues related to assigning gestures to text and converting the text into agent animations synchronized with speech. First, our empirical study identified useful lexical and syntactic information for assigning gestures to plain text. Specifically, when a bunsetsu unit is a constituent of coordination, gestures occur almost half the time. Gestures also frequently co-occur with nominal phrases modified by a clause. These findings suggest that syntactic structure is a stronger determinant of gesture occurrence than theme or rheme and given or new information specified by local grammatical cues.

We plan to enhance our model by incorporating more general discourse level information, though the current system exploits cue words as a very partial kind of discourse information. For instance, gestures frequently occur at episode boundaries. Pushing and popping of a discourse segment (Grosz & Sidner, 1986) may also affect gesture occurrence. Therefore, by integrating a discourse analyzer into the LTM, more general structural discourse information can be used in the model. Another important direction is to evaluate the effectiveness of agent gestures in actual human-agent interaction. We expect that if our model can generate gestures with appropriate timing for emphasizing important words and phrases, users can perceive agent presentations as being more alive and comprehensible. We plan to conduct a user study to examine this hypothesis.

References

Andre, E., Rist, T., & Muller, J. (1999). Employing AI methods to control the behavior of animated interface agents. *Applied Artificial Intelligence, 13*, 415-448.

Berger, K. W., & Popelka, G. R. (1971). Extra-facial Gestures in Relation to Speech-reading. *Journal of Communication Disorders, 3*, 302-308.

Cassell, J. et al. (1994). *Animated Conversation: Rule-Based Generation of Facial Expression, Gesture and Spoken Intonation for Multiple Conversational Agents.* Paper presented at the SIGGRAPH '94.

Cassell, J., & Prevost, S. (1996). *Distribution of Semantic Features Across Speech and Gesture by Humans and Computers.* Paper presented at the Workshop on the Integration of Gesture in Language and Speech.

Cassell, J., Stone, M., & Yan, H. (2000). *Coordination and Context-Dependence in the Generation of Embodied Conversation.* Paper presented at the INLG 2000.

Cassell, J., Vilhjalmsson, H., & Bickmore, T. (2001). *BEAT: The Behavior Expression Animation Toolkit.* Paper presented at the SIGGRAPH 01.

Firbas, J. (1971). On the Concept of Communicative Dynamism in the Theory of Functional Sentence Perspective. *Philologica Pragensia, 8*, 135-144.

Givon, T. (1985). Iconicity, Isomorphism and Non-arbitrary Coding in Syntax. In J. Haiman (Ed.), *Iconicity in Syntax* (pp. 187-219): John Benjamins.

Grosz, B., & Sidner, C. (1986). Attention, Intentions, and the Structure of Discourse. *Computational Linguistics, 12*(3), 175-204.

Halliday, M. A. K. (1967). *Intonation and Grammar in British English*. The Hague: Mouton.

Kendon, A. (1972). Some Relationships between Body Motion and Speech. In A. W. Siegman & B. Pope (Eds.), *Studies in Dyadic Communication* (pp. 177-210). Elmsford, NY: Pergamon Press.

Kurohashi, S., & Nagao, M. (1994). A Syntactic Analysis Method of Long Japanese Sentences Based on the Detection of Conjunctive Structures. *Computational Linguistics, 20*(4), 507-534.

Lester, J. C., Stone, B., & Stelling, G. (1999). Lifelike Pedagogical agents for Mixed-Initiative Problem Solving in Constructivist Learning Environments. *User Modeling and User-Adapted Interaction, 9*(1-2), 1-44.

McNeill, D. (1992). *Hand and Mind: What Gestures Reveal about Thought*. Chicago, IL/London, UK: The University of Chicago Press.

Prevost, S. A. (1996). *An Informational Structural Approach to Spoken Language Generation.* Paper presented at the 34th Annual Meeting of the Association for Computational Linguistics, Santa Cruz, CA.

Rickel, J., & Johnson, W. L. (1999). Animated Agents for Procedural Training in Virtual Reality: Perception, Cognition and Motor Control. *Applied Artificial Intelligence, 13*(4-5), 343-382.

Rogers, W. (1978). The Contribution of Kinesic Illustrators towards the Comprehension of Verbal Behavior within Utterances. *Human Communication Research, 5*, 54-62.