3D1-07

Gestures Realization for Embodied Conversational Agents

Qing LI[*1] Yukiko NAKANO[*2] Toyoaki NISHIDA [*1]

[*1] Graduate school of Information Science and Technology, the University of Tokyo

[*2] Research Institute of Science and Technology for Society

Embodied Conversational Agent (ECA) is a life-like virtual human capable of carrying on conversations with humans by both understanding and producing verbal and nonverbal behaviors. In human's conversation people make gestures while speaking, which can help them express themselves clearly. To be a means of human-computer interaction (HCI), it is essential for ECA to realize the automatic synchronization of gesture with speech. In this paper, we take an analysis of gestures employed by lecturers in symposiums and propose rules we obtained to our presentation agent system. Our work improves a basic understanding of integration of speech with nonverbal behaviors for building ECA. Gestures synthesized by our system are expected to be convincing, and are more natural after all the rules proposed are implemented.

1. Introduction

ECAs are computer interfaces that have bodies and know how to use them for conversation. ECAs are specifically humanlike in the way they use their bodies in conversation with the ability to generate verbal and nonverbal output. Due to the development in 2D and 3D computer animation, the performance of animated characters becomes more and more natural. Synchronization of gesture with speech output attracts considerable attention from HCI designers and becomes a new paradigm for the research on ECAs.

In our research we select presentation as our application domain because presentation is an effective way for people to forward their information to many people. There are a lot of researches on multimodal presentation agents because multimodal agents make presentations more attractive and more persuasive. Most of researches like [André 1998] provide a script language allowing users to design presentations, which still has a problem that users should learn the script language in advance. The motivation to our research is to relieve the user's burden from designing a multimodal presentation by providing a presentation agent which can generate nonverbal behaviors especially gesture automatically from the input text alone. Besides, we try to find some insight into the relationship between nonverbal behaviors and the discourse structure of spoken language, which may provide a basis of building ECA system.

The rest of this paper is organized as follows. In the second section, related work on speech-related gesture is introduced. The third section introduces our agent system CAST (the Conversational Agent System for neTwork application) [Nakano 2003]. Then, we analyze gestures, summarize rules of the relation among gesture, gaze and spoken language, and propose rules for CAST. Finally, the conclusion and the future work are shown.

2. Related Work

[McNeill 1992] classified gesture movements into four major categories: Iconic, Metaphoric, Deictic (pointing), and Beat gestures. Iconic gestures bear a close formal relationship to the semantic content of speech. Metaphoric gestures are similar to Iconic gestures in that they present imagery, but present an image of an abstract concept. Deictic gestures are pointing movements. Beat gestures are formless gestures without an independent meaning.

Gesture-Unit is defined as the period of time between successive resets of the limbs. It begins the moment the limb begins to move and ends when it has reached a rest position again. Gesture-phrase is a unit of one gesture. And Gesture-Phrase consists of one or more movement phases (preparation, various holds, stroke, and retraction).

- Preparation (optional), in which the limb moves away from its rest position to a position in gesture space where the stroke begins.
- Pre-stroke hold (optional), is the position and hand posture reached at the end of the preparation itself. A hold in general is any temporary cessation of movement without leaving the gesture hierarchy.
- Stroke (obligatory) is the peak of effort in the gesture.
- Post-stroke hold (optional) is the final position and posture of the hand reached at the end of the stroke; this may be held more or less briefly until the retraction begins.
- Retraction (optional) is the return of the hand to a rest position.

Previous work found that indexical gestures are likely to co-occur with the focused part of a spoken sentence, and that the stroke onset co-occurs with the most contrastive words or phrases in speech and co-varies with it in time. Despite these findings it is still far from implementing such rules in application of multimodal agent system. So we analyze the video data to understand when, where and how people gesture to get a set of rules on the relation between gesture and speech to fill the gap.

3. Agent system

Our agent system (CAST) selects appropriate nonverbal behaviors according to the linguistic information in the text, and

Contact: Li Qing
Nishida & Kurohashi Lab
Graduate School of Information Science and Technology
The University of Tokyo
7-3-1 Hongo, Bunkyo-ku, Tokyo 113-8656, Japan
Phone & fax: +81-3-5841-6689
E-mail: liqing@kc.t.u-tokyo.ac.jp

generates character animation synchronized with synthesized speech. The architecture of CAST is shown in Figure 1. It is composed of four main modules: the Agent Behavior Selection Module (ABS), the Language Tagging Module (LTM), a character animation system, and a Text-to-Speech engine (TTS).

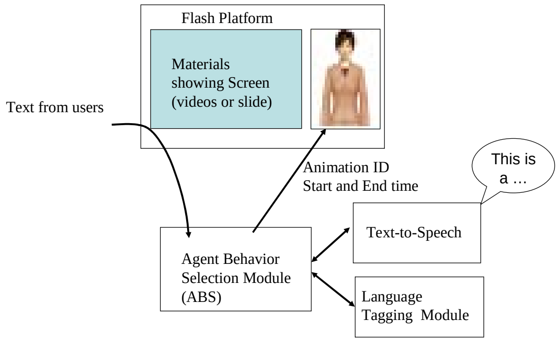

Figure 1 the architecture of CAST

The work flow is: (1) LTM tags a text input by annotating with syntactic information; at the same time, TTS calculates the timing information of the input text; (2) ABS selects appropriate gestures and facial expressions using linguistic information obtained from the LTM; (3) ABS calculates a time schedule for each agent action according to timing information from TTS; (4) finally ABS starts the animation action in animated system and speech audio in TTS in a synchronized way.

From the description of the workflow, it is obvious that ABS is the principal component in the agent system. We use BEAT (The Behavior Expression Animation Toolkit) [Cassell 2001] as a basis of the ABS. The nonverbal behaviors generation module of ABS is to select appropriate nonverbal behaviors using linguistic and contextual information contained in the typed text; the behaviors scheduling module of ABS is to calculate a time schedule in order to execute the behaviors and speech in a synchronized way.

In our research, we use a natural language processing tool for Japanese [Kurohashi 1994] to implement the Language Tagging Module (LTM) and a Flash-based character animation system (RISA) to provide the appearance and actions.

4. Gesture analysis

4.1 Data and Annotation

(1) Data

We select about 5-minute videos from five people's presentation in one symposium. The symposium was held in a big meeting room, where there is a big screen and a desk for lecturers before the audience's seats. Among these five people, one lecturer uses slides, and the other use PPT files. All of them did not know they were taken to analyze their gesture. And the topics these lectures talk about are wide from social science research to engineering research. It is worthy to be mentioned that all lecturer holding a microphone, which may decreases the number of gesture and prevents them from gesture with two hands. But people gesture during presentation even with one hand free. It is the naturalness that makes us to choose these presentation videos.

(2) Annotation

We use a generic video annotation tool Anvil [Kipp 2003] to annotate our data. Anvil offers frame-accurate, hierarchical multi-layered annotation with objects that contain attribute-value pairs. Layers and attributes are all user-defined. Its coding takes place on a time-aligned annotation board that can be customized with color-coding to allow efficient and intuitive annotation. Utilizing Anvil, we define tracks including Slide, Face, Wave, Discourse structure and Gesture (Unit, Phase, and Phrase).

- "Slide" track tells which slide the lecturer is using.
- "Face" track is to annotate the gaze of lecturer. In our research there are three values: facing note computer or notes paper, facing audience, and facing screen.
- "Wave" is the track to import audio wave.
- "Transcription" track shows the transcription of utterance.
- "Gesture" group consists of Gesture-"Unit", Gesture-"Phrase" and Gesture-"Phase".
- "Discourse structure" track is to annotate the discourse structure.

The reason why we annotate the slide and the gaze is that we suppose that there are clues among gaze and language and gesture. We are also interested in the timing and planning among the materials lecturer use, verbal behaviors and nonverbal behaviors to apply the rules to multimodal presentation agent system.

4.2 Gesture classification

The classification for gesture is based on [McNeill 1992] in our research. Taking the discourse structure into consideration, we classify gestures into Iconic gesture, Metaphoric gesture, deictic gestures, Beat gestures, and Contrast gestures. We subcategorize beat gesture into three groups according to their different features in gesture space and gesture phase, which facilitates finding insight into the relation between beat gesture and language.

- Beat-1: Beat gesture with a post-hold (phases: stroke and then hold)
- Beat-2: Beat gesture without post-stroke hold
- Beat-3: In the successive beat gestures, they are gestured in different gesture space.

According to the above definition, when gestures are coded, three factors should be considered in order. They are 1) shape of gesture, 2) the timing, 3) semantic relation. The order we utilize is 1)→2)→3) because the shape of gesture is the most objective factor. Based on this coding process, we code the gesture types by eyes.

4.3 Results

(1) General results

It is often said that Japanese people seldom use gestures when they give presentations. In the videos we took, over half of people use few gestures. But almost everyone use gestures when he answers the question asked by the audience. The possible explanation is that people tend to make more gestures when they speak while thinking. We found that presentations with gesture show more attractive than presentations without gestures, which strengthens our belief that a presentation agent able to make gestures automatically and naturally can give a presentation on the behalf of people.

Table 1 Distribution of Gesture Unit over all people

G-Unit	Number	Percent (%)
1 gesture	99	57.23
2 gestures	36	20.81
3 gestures	10	5.87
4 gestures	11	6.36
5 gestures	6	3.47
More than 5 gestures	11	6.36
Total	173	------

After analyzing video data, we obtained general results about the gestures distribution. It is found that these five people make a gesture per 5.52second on the average. Table 1 indicates that the distribution of the Gesture-Unit. The cases of one gesture and two gestures in one Gesture-Unit amount to about 80% in the number of all Gesture-Unit together. These findings may provide us useful information to implement how often presentation agents should make a gesture.

Table 2 Gesture distribution over all five people

Figure Gesture type	number	Percent (%)
Iconic gesture	0	0
Metaphoric gesture	17	7.05
Deictic gesture	44	18.26 *
Beat gesture	134	55.60 *
Contrast gesture	2	0.83
Uncertain gesture	44	18.26
(Total)	241	-----

Table 2 shows that the two most used types of gestures are beat gestures and deictic gestures with the proportion to the total number of gestures of 55.6%, 18.26%, respectively. And the iconic and metaphoric gestures are each less than 8%. The possible reason for the increase in using deictic gesture is that the data videos where people give presentation with slides. In the following, we will analyze the relation between beat gesture and discourse structure, and the deictic gesture and the content of slide further. In addition, we summarize rules on gaze.

(2) Understanding beat gesture

It is reported [McNeill 1992] that the stroke (energetic and meaningful) is synchronized with the linguistic segments that are co-expressive with it. So we relate the words within the stroke and gesture. We find that beat gestures are likely to occur in the following cases in the view of the discourse structure:

- Case1: Important words following conversion markers like "not …but "*(DEWANAKUTE).*
- Case2: Important words following conclusion and paraphrase markers like "in a word" *(YOUSURUNI)* and "in other words"*(TSUMARI).*
- Case3: Each of enumerated words or phrases.
- Case4: Some adverbs for emphasis, like " very " *(HIJOUNI),* "originally"*(HONNRAI).*
- Case5: Interrogatives except for those in a nominal clause.
- Case6: Exemplified items after an example marker.

Table 3 the possibility of using beat gestures

Language information	The number of all cases	Using gesture	Using case / all case
Case1	10	7	70%
Case2	7	5	71%
Case3	12	9	75%
Case4	23	18	78%
Case5	14	10	71%
Case6	10	7	70%

Table 3 indicates the possibility of using beat gestures in the above cases. Based on this table, we will propose suggestion to beat generator by utilizing JUMAN [Kurohashi 2000] and KNP [Kurohashi 2000]. When these rules are applied, one point needs to be considered especially. It is that how to detect words for beat gestures should depend not only on key word matching but also on discourse structure. For example, when a conclusion marker is detected, it doesn't mean put the stroke of beat gesture on this marker but on the following important words, which is still a challenge for natural language processing research. So we suggest rules to beat generator only in case3, case4 and case5 which can be realized by key word matching. The rest other implementations is our future work.

(3) Relation between the content of slides and deictic gesture

There are two types of deictic gestures in presentation: (a) one is used with touching (or very close to) the screen; (b) the other type is used quite far from the screen. The difference is that the type (a) is used to point at an exact position on the screen. On the other hand, type (b) can only indicate the direction of pointing (ie. pointing towards the screen). In presentation, type (b) deictic gesture is frequently used. Thus, we focus on type (b) in our deictic gesture analysis. We analyze phrases which are accompanied by a deictic gesture. Table 5 and Table 6 show the distribution of deictic gesture by considering the syntactic type of phrase and type of referent, respectively.

Table 4 Distribution of deictic gesture based on syntactic type

Syntactic type of phrase	Percentage
Demonstratives	47%
object name on the slide	29%
other	24%

Table 5 Distribution of deictic gesture considering referent type

Type of referent	Percentage
a whole one slide, or part of a slide	65%
an object on a slide	18%
words or phrases on a slide	6%

In addition, when a demonstrative whose referent cannot be resolved using the linguistic context, a deictic gesture is used 57% of the time. These results indicates that deictic gestures are most frequently used with demonstratives, and they more frequently refer to a region of a slide that is not referred in the previous discourse.

(4) Gaze

When a presenter talks about information on the slide, s/he sometimes looks at the slide and sometimes doesn't. For example, s/he reads a whole one line on a slide, or utters a sentence including words shown on a slide. We analyzed what type of information on a slide is presented with looking at the slide. In presenter's speech, we picked out words and phrases shown on a slide, and classified them according to the place they appear on the slide. The categories are:(1)In Title;(2)In top level bullet(3)In a bullet lower than the top level;(4)In table or graph.

We find that when words or phrases are included in a title of a slide, a presenter looks at the slide (like reading them) 57% of the time. The bullet is getting smaller; the proportion of looking at the slide is getting higher. This result suggests that a presenter more frequently looks at the slide when s/he talks about details.

(5) Rules for Nonverbal Generator of our agent system

We propose three models on beat gesture, deictic gesture and gaze as follows:

```
<Beat>
IF     Each of enumerated words or phrases
       OR Adverbs for emphasis
       OR Interrogatives except for those in a nominal
          clause are detected
THEN   Use Beat gestures 70% of the cases
<Deictic>
IF   a phrase includes demonstratives, especially "this" or "these",
   IF    the referent is found in the previous discourse,
         THEN   Do not use deictic
   ELSE
         THEN   Use deictic 57% of the time
<Interaction between eye gaze and deictic gesture>
IF   a deictic gesture is selected,
       THEN   Look at the screen 82% of the time
<Gaze>
IF    a phrase is included in a title,
       THEN   Look at the screen 57% of the time
ELSEIF a phrase is included in a level 1 bullet,
       THEN   Look at the screen 71% of the time
ELSEIF a phrase is included in a bullet lower than level 1,
       THEN   Look at the screen
ELSEIF a phrase or number is included in a graph or table,
       THEN   Look at the screen 94% of the time
```

5. Discussion and Future work

We analyzed nonverbal behaviors of the people who give presentation on symposiums. The focus of our research is put on gestures e specially on beat gestures and deictic gestures because they are the two most used types among gestures used in presentations. Based on the research of speech-related gesture, we summarized the relation between beat gestures and the discourse structure, and the relation between deictic gestures and the content of the slide, as well as between gaze and the content of the slide. We obtained rules of the co-occurrence of these nonverbal behaviors with speech and proposed rules for our agent system CAST which takes plain text as input, and automatically generates a presentation featured with an animated agent. Gestures generated by CAST are expected convincing and more natural after all the rules proposed here are implemented.

In spite of the fact that we have found some rules on gestures and discourse structure by now, our present work is just a beginning for our goal of realizing nonverbal behaviors automatic generation for a presentation agent. Analysis work is the base which makes the agent system possible to behave like a human being, so the analysis work is still one of focuses of our research. As we know, when a nonverbal behavior is implemented in a wrong place, the naturalness of ECA will be crashed completely. We need analyze more data to clarify some ambiguous rules. Moreover, it seems that different kinds of nonverbal behaviors like gaze, facial expression interact with each other. For example, people often look at the audience to make a metaphoric gesture and make a deictic gesture after he turns to the direction. Thus, we will analyze not only various kinds of nonverbal behaviors but also the relation among them to provide knowledge of timing and scheduling nonverbal behaviors for building ECAs.

Acknowledgement

Our warmest thanks to Natural Language Processing Group in Nishida & Kurohashi Lab, especially to Masashi OKAMOTO, Ph.D., who gives much advice on discourse structure and assists our analysis work.

References

[Andre 1998] Elisabeth André, Thomas Rist, Jochen Müller: Integrating Reactive and Scripted Behaviors in a Life-Like Presentation Agent, in: Proc. of the Second International Conference on Autonomous Agents (Agents '98), pp. 261-268, 1998.

[Cassell 2001] J. Cassell, H. Vilhjalmsson, and T. Bickmore. BEAT: the Behavior Expression Animation Toolkit. In Proceedings of SIGGRAPH-01, pp. 477--486, 2001.

[Kurohashi 1994] Kurohashi, S. and M. Nagao, A Syntactic Analysis Method of Long Japanese Sentences Based on the Detection of Conjunctive Structures. Computational Linuguistics,) p. 507-534. 1994. 20(4).

[Kurohashi 2000] Sadao Kurohashi, Nagao Makoto, Japanese Morphological Analysis System JUMAN Version 3.6.2, Graduate School of Informatics of Kyoto University, 2000.

[Kurohashi 2000] Sadao Kurohashi, Japanese Parser KNP Version2.0 Manual, Graduate School of Informatics of Kyoto University, 2000.

[Kipp 2003] Michael Kipp, Anvil 4.0 Manual, 2003.

[McNeill 1992] D. McNeill, Hand and Mind: What gestures Reveal about Thought. University of Chicago Press, Chicago, 1992.

[Nakano 2003]Yukiko I. Nakano, Toshihiro Murayama, Daisuke Kawahara, Sadao Kurohashi, and Toyoaki Nishida "Embodied Conversational Agents for Presenting Intellectual Multimedia Contents" Seventh International Conference on Knowledge-Based Intelligent Information & Engineering Systems (KES'2003), September 3-5, 2003, University of Oxford, United Kingdom (to appear).