Towards Generation of Fluent Referring Action in Multimodal Situations

Tsuneaki Kato
NTT Information and
Communication Systems Labs.
Yokosuka, Kanagawa 239, JAPAN
kato@nttnly.isl.ntt.co.jp

Yukiko I. Nakano
NTT Information and
Communication Systems Labs.
Yokosuka, Kanagawa 239, JAPAN
yukiko@nttnly.isl.ntt.co.jp

Abstract

Referring actions in multimodal situations can be thought of as linguistic expressions well coordinated with several physical actions. In this paper, what patterns of linguistic expressions are commonly used and how physical actions are temporally coordinated to them are reported based on corpus examinations. In particular, by categorizing objects according to two features, visibility and membership, the schematic patterns of referring expressions are derived. The difference between the occurrence frequencies of those patterns in a multimodal situation and a spoken-mode situation explains the findings of our previous research. Implementation based on these results is on going.

1 Introduction

A lot of active studies have been conducted on the temporal coordination of natural language and visual information. The visual information considered includes pointing gestures (André & Rist, 1996), facial expressions and iconic gestures (Cassell et al., 1994), and graphical effects such as highlighting and blinking (Dalal et al., 1996; Feiner et al., 1993). Among those we have been focusing on generating effective explanations by using natural language temporally coordinated with pictures and gestures. The experimental system we implemented is for explaining the installation and operation of a telephone with an answering machine feature, and simulates instruction dialogues performed by an expert in a face-to-face situation with a telephone in front of her (Kato et al., 1996). The system explains by using synthesized speech coordinated with pointing gestures from a caricatured agent and simulated operations implemented by the switching of figures. One of the important issues for enhancing this type of system is to shed light on what makes referring actions fluent in multimodal situations and to build a mechanism to generate such fluent actions.

We also empirically investigated how communicative modes influence the content and style of referring actions made in dialogues (Kato & Nakano, 1995). Experiments were conducted to obtain a corpus consisting of human-to-human instruction dialogues on telephone installation in two settings. One is a spoken-mode dialogue situation (SMD hereafter), in which explanations are given using just voice. The other is a multimodal dialogue situation (MMD hereafter), in which both voice and visual information, mainly the current state and outlook of the expert's telephone and her pointing gestures to it, can be communicated. Detailed analysis of the referring actions observed in that corpus revealed the following two properties.

P1: The availability of pointing, communication through the visual channel reduces the amount of information conveyed through the speech or linguistic channel. In initial identification, the usage of linguistic expressions on shape/size, characters/marks, and related objects decreases in MMD, while the usage of position information does not decrease.

P2: In SMD, referring actions tend to be realized to an explicit goal and divided into a series of fine-grained steps. The participants try to achieve them step by step with many confirmations.

Although our findings were very suggestive for analyzing the properties of referring actions in multimodal situations, they were still descriptive and not sufficient to allow their use in designing referring action generation mechanisms. Then, as the next step, we have been examining that corpus closer and trying to derive some schemata of referring actions, which would be useful for implementation of multimodal dialogue systems. This paper reports the results of these activities.

Two short comments must be made to make our research standpoint clearer. First, our purpose is to generate referring actions that model human referring actions in mundane situations. Theoretically speaking, as Appelt pointed out, it is enough for referring to provide sufficient description to distin-

guish one object from the other candidates (Appelt, 1985). For example, a pointing action to the object must be enough, or description of the object's position, such as "the upper left button of the dial buttons" also must be considered sufficient. However, we often observe referring actions that consist of a linguistic expression, "a small button with the mark of a handset above and to the left of the dial buttons", accompanied with a pointing gesture. Such a referring action is familiar to us even though it is redundant from a theoretical viewpoint. Such familiar actions that the recipient does not perceive as awkward is called fluent in this paper. Our objective is to generate such fluent referring actions, and is rather different from those of (Appelt, 1985) and (Dale & Haddock, 1991).

Second, in our research, a referring action is considered as the entire sequence of actions needed for allowing the addressee to identify the intended object and incorporating its achievement into part of the participants' shared knowledge. In order to refer to an object in a box, an imperative sentence such as "Open the box, and look inside" may be used. Such a request shifts the addressee's attention, and to see it as a part of the referring action may be problematic. It is, however, reasonable to think that both the request for looking into the box and the assertion of the fact that an object is in the box come from different plans for achieving the same goal, identifying the object. As Cohen claimed that it is useful to understand referring expressions from the viewpoint of speech act planning (Cohen, 1984), it is not so ridiculous to go one step further and to consider the entire sequence of actions, including attention shifts, as an instance of a plan for object referring. Moreover, this approach better suits implementing a referring action generation mechanism as a planner.

The next section describes what kinds of linguistic expression are used for referring actions in MMD and compares them with those in SMD. In particular, by categorizing objects according to two features: visibility and membership, schemata for object referring expressions of each category are derived. In the third section, how several kinds of actions such as pointing gestures are accompanied by such expressions is reported. In the fourth section, implementation of referring action generation is discussed based on our findings described thus far. Finally, in the last section, our findings are summarized and future work is discussed.

2 Linguistic expression in referring actions

Referring actions in multimodal situations can be thought of as linguistic expressions well coordinated with several physical actions. The linguistic expressions for referring to objects, referring expressions, are focused on in this section, and in the next section, how those expressions should be coordinated with actions is discussed.

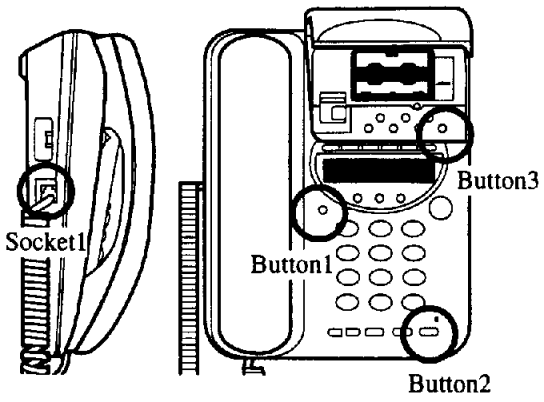

Figure 1: The telephone used in the corpus

2.1 Object categorization

The top and left side views of the telephone used are shown in Fig. 1. Although the objects such as buttons can be identified by using several features such as position, color, shape, and size, the two features described below proved to be dominant in the referring expressions used.

Visibility: Some objects are located on the side or back of the telephone, and can not be seen unless the body is turned over or looked into. Some objects lie underneath the cover, and opening that cover is needed in order to see them. Such objects are categorized into invisible ones and distinguished from visible ones, which are located on the top[1].

Membership: Aligned buttons of the same shape and color are usually recognized as a group. Members of such a group are distinguished from isolated ones[2].

In Fig. 1, socket 1 on the side is invisible and isolated, button 1 on the left of the top surface is visible and isolated, button 2 on the lower right of the top surface is a visible group member, and button 3 on the upper right is an invisible group member as it is underneath a cassette cover usually closed.

According to this categorization, we determined which patterns of referring expressions were frequently observed for each type of object. The patterns thus extracted can be expected to yield

[1] As you have already realized, this feature is not intrinsic to the object, but depends on the state of the telephone when the object is referred to. Buttons underneath the cover are visible when the cover is open.

[2] The recognition of a group may differ among people. In daily life, however, we believe an effective level of consensus can be attained.

schemata for referring expression generation. Three explanations of five experts in two situations, MMD and SMD, i.e. fifteen explanations in each situation, were analyzed. The apprentices differed with each explanation. Every referring expression analyzed involved initial identification, which is used to make the first effective reference to an object, and to introduce it into the explanation. All objects were referred to in the context in which the expert made the apprentice identify it and then requested that some action be performed on it. All explanations were done in Japanese[3].

2.2 Schemata for referring to visible isolated objects

Referring actions to visible isolated objects are rather simple and were basic for all cases. Two major patterns were observed and can be treated as the schemata. Table 1 shows these two schemata[4], called *RS1* and *RS2* hereafter. *RS1* contains two sentences. The first asserts the existence of the object at a described position. The second is an imperative sentence for requesting that an action be performed on the object identified. In the first sentence, a postpositional phrase describing the object position precedes the object description. The object description is a noun phrase that has modifiers describing features of the object such as its color or size followed by a head common noun describing the object category. That is, its structure is

[object description *np*] →
[feature description *pp/comp*] *
[object class name *n*]

In *RS2*, the imperative sentence requesting an action contains a description referring to the object. This description has the same structure as *RS1* shown above. In most cases, the first feature description is a description of the object position. In both schemata, object position is conveyed first, other features second, and the requested action follows. This order of information flow seems natural for identifying the object and then acting on it. Examples of referring expressions[5] that fit these schemata are

(1) daiarubotan no hidariue ni juwaki no
dial-buttons upper-left LOC handset
maaku ga tsuita chiisai botan
mark SUBJ being-placed-on small button
ga arimasu. sore wo oshi tekudasai.
SUBJ exist. it OBJ push REQUEST
'On the upper left of the dial buttons, there is a small button with the mark of a handset. Please push it.'

(2) daiarubotan no hidariue no juwaki no maaku
dial-buttons upper-left handset mark
ga tsuita chiisai botan wo oshi
SUBJ being-placed-on small button OBJ push
tekudasai.
REQUEST
'Please push the small button with the mark of a handset on the upper left of the dial buttons.'

In *RS1*, the achievement of identification is confirmed by making the first sentence a tag question or putting a phrase for confirmation after that sentence. Sometimes it is implicitly confirmed by asserting the existence of the object as the speaker's belief. In *RS2*, confirmation can be made by putting a phrase for confirmation after the noun phrase describing the object or by putting there a pause and a demonstrative pronoun appositively.

Another pattern was observed in which *RS1* was preceded by an utterance referring to a landmark used in the position description. This is also shown in Table 1 as *RS11*. In *RS11*, reference to the landmark is realized by an imperative sentence that directs attention to the landmark or a tag question that confirms its existence. Examples are

(3) hontai no hidariue wo mi tekudasai. soko
body upper-left OBJ look REQUEST there
ni chiisai botan ga arimasu.
LOC small button SUBJ exist
'Please look at the upper left of the body. There is a small button there.'

(4) daiarubotan no 1 arimasu yone. sono
dial-button 1 exist CONFIRM its
hidariue ni chiisai botan ga arimasu.
upper-left LOC small button SUBJ exist
'There is dial button 1, isn't there? On its upper left, there is a small button.'

Table 1 shows the numbers of occurrences of each pattern in MMD and SMD. The total occurrence number was 30, as two objects fell under this category. *RS11* and *RS1* frequently occur in SMD.

[3] Japanese is a head-final language. Complements and postpositional phrases on the left modify nouns or noun phrases on the right, and construct noun phrases. That is, a simplified version of Japanese grammar contains $pp \rightarrow np\ p$, $np \rightarrow pp\ np$, $np \rightarrow comp\ np$, and $np \rightarrow n$. Sentences are constructed by a rule, $s \rightarrow pp* v$. The order of *pp*s is almost free syntactically, being determined by pragmatic constraints. Older information precedes newer (Kuno, 1978).

[4] Schemata are represented as sequences of terminal symbols, non terminal symbols each of those has a form of [semantic content *syntactic category*], and *schema ID*. A slash appearing in a syntactic category means options rather than a slash feature.

[5] All examples are basically extracted from the corpus examined. Those, however, were slightly modified by removing non-fluencies and the diversities caused by the factors mentioned in section 2.4 below.

Table 1: The schemata for referring to visible isolated objects and their occurrence frequency

ID	Pattern/Descritpion	MMD	SMD
RS1	[position np] ni(LOC) [object description np] ga(SUBJ) arimasu(*exist*). [object np] wo(OBJ) [action v] tekudasai(REQUEST).	12	19
RS2	[object description np] wo(OBJ) [action v] tekudasai(REQUEST)	13	5
RS11	[referring to a landmark s], *RS1*	0	4
Others		5	2

2.3 Schemata for referring to invisible objects and group members

Five objects fell into the category of invisible isolated objects. Two schemata described in the previous subsection, *RS1* and *RS2*, were used for referring to these objects by conveying the fact of which surface the object was located on as the position description. For example,

(5) hontai no hidarigawa ni sashikomiguchi ga
body left-side LOC socket SUBJ
arimasu. soko ni sore wo ire tekudasai.
exist there LOC it OBJ put REQUEST
'There is a socket on the left side of the body. Please put it there.'

(6) sore wo hontai hidarigawa no sashikomiguchi
it OBJ body left-side socket
ni ire tekudasai
LOC put REQUEST
'Please put it into the socket on the left side of the body.'

In addition, *RS11* and its *RS2* corespondent, *RS12*, were used frequently. In these patterns, the surface on which the object is located is referred to in advance. It is achieved by an imperative sentence that directs attention to the surface or asks that the body be turned, or by an description of the side followed by a confirmation. Examples are

(7) hontai hidarigawa no sokumen wo mi
body left side OBJ look
tekudasai. soko ni ...
REQUEST there LOC ...
'Please look at the left side of the body. On that side, ...'

(8) hontai no hidari sokumen desu ne.
body left side COPULA CONFIRM
soko no mannaka ni ...
there center LOC ...
'The left side of the body, you see? On the center of that side, ...'

Table 2 shows the schemata based on these patterns and their numbers of occurrence; the total is 75. *RS2* is frequently used in MMD, while *RS11* is frequently used in SMD.

For referring to a visible group member, patterns are observed in which the group the object belongs to is referred to as a whole, in advance, and then the object is referred to as a member of that group. The first sentence of *RS1* is mainly used for referring to the group as a whole. For example,

(9) daiarubotan no shita ni onaji iro no
dial-buttons below LOC SAME color
botan ga itsutsu narande imasu.
buttons SUBJ five aligned be
'Below the dial buttons, there are five buttons of the same color.'

After this, *RS1* or *RS2* follows. These patterns, hereafter called *RS21* and *RS22*, respectively, are shown in Table 3. In each pattern, the relative position of the object in the group is used as the position information conveyed later. In *RS21*, the following sentence, for example, follows the above.

(10) sono migihashi ni supiika no maaku ga
those right-most LOC speaker mark SUBJ
tsuita botan ga arimasu
being-placed-on button SUBJ exist
'On the right most of those, there is a button with the mark of a speaker.'

RS1 and *RS2*, in which a referring expression to a group does not constitute an utterance by itself are also observed, such as

(11) ichiban-shita no botan no retsu no migihashi
bottom buttons line right-most
ni supiika no maaku ga tsuita
LOC speaker mark SUBJ being-placed-on
botan ga arimasu.
button SUBJ exist
'On the right most of the line of buttons on the bottom, there is a button with a mark of a speaker.'

In the above, although the expression referring to the group is part of the expression referring to the

Table 2: The schemata for referring to invisible objects and their occurrence frequency

ID	Pattern/Descritpion	MMD	SMD
RS1		16	11
RS2		23	7
RS11	[referring to the side s/np], *RS1*	10	33
RS12	[referring to the side s/np], *RS2*	5	5
Others		21	19

member, information that the object is a member of a specific group is conveyed and the position relative to the group is used for describing the object's position. There are other patterns which do not contain such descriptions of groups at all. For example,

(12) hontai migishita no supiika botan wo oshi
body right-lower speaker button OBJ push
tekudasai.
REQUEST
''Push the speaker button on the lower right of the body.'

According to this characteristic, *RS1* and *RS2* are divided into two patterns. *RS1* and *RS2* with descriptions of a group are called *RS1'* and *RS2'* respectively, and *RS1* and *RS2* without descriptions of a group are called *RS1"* and *RS2"*. Table 3 shows the numbers of occurrence. The total number is 60, as four objects fell into this category[6]. *RS1"* and *RS2"* are frequently observed in MMD, while *RS21* and *RS22* are frequently observed in SMD.

Just one object was an invisible group member in our corpus. It was the button underneath the cassette cover. All referring expressions in both MMD and SMD contain an imperative sentence requesting that the cassette cover be opened. It is considered that this imperative sentence corresponds to the imperative sentences in *RS11* and *RS12* that direct attention to the side of the body or ask that the body be turned. Subsequent referring expressions follow the same patterns as for visible group members. The distribution of the patterns is also similar. That is, the schemata for referring to invisible group members are obtained as combinations of those for invisible objects and group members.

2.4 Factors that complicate referring expressions

The previous two subsections derived the schemata for referring expressions in line with the objects' categorization based on two features. The schemata are just skeletons, and referring expressions with more diverse forms appear in the collected corpus. The most important origin of this diversity is that explanation dialogue is a cooperative process (Clark & Wilkes-Gibbs, 1990). First, several stages of a referring action can trigger confirmation. Those confirmations are realized by using various linguistic devices such as interrogative sentences, tag questions, and specific intonations. Second, related to incremental elaboration, appositive and supplemental expressions are observed. For example,

(13) rusu botan arimasu ne, gamen no
OUT button exist CONFIRM display
shita, "rusu" to kakareta shiroi botan.
under "OUT" with being-labeled white button
'There is an OUT button, under the display, a white button labeled "OUT."'

These inherent dialogue features complicate referring expressions. Moreover, it is difficult to derive patterns from exchanges in which the apprentice plays a more active role such as talking about or checking her idea on the procedure in advance.

The second origin of diversity relates to the fact that experts sometimes try to achieve multiple goals at the same time. Labeling an object with a proper name is sometimes achieved simultaneously with identifying it. This phenomena, however, could be schematized to some extent. Two patterns are observed. The one is to put the labeling sentence such as "This is called the speaker button" after the first sentence in *RS1* or the noun phrase describing the object in *RS2*. The other is to use a proper name as the head of the noun phrase describing the object. An example is "the speaker button with the mark of a speaker".

The third origin is the effect of the dialogue context which is determined external to the referring expressions. For example, almost half of the referring expressions categorized into *Others* in the above tables fit one of the following two patterns, called *RS3* hereafter.

[object function $pp/comp$] [object np] ga(SUBJ)
[position np] ni(LOC) arimasu(*exist*).
[description of the features of the object s] *

[6] One object belonged to a group that contained an object already referred to. This implies that the group had already been identified. The usage of *RS21* and *RS22* was relatively scarce for that object. This suggests that referring expressions should be affected by the history of the group as well as of the object itself.

Table 3: The schemata for referring to group members and their occurrence frequency

ID	Pattern/Descritpion	MMD	SMD
RS21	[referring to the group $_s$], *RS1*	4	12
RS22	[referring to the group $_s$], *RS2*	7	15
RS1'	*RS1* (with group descriptions)	0	4
RS2'	*RS2* (with group descriptions)	7	9
RS1"	*RS1* (w/o group descriptions)	12	7
RS2"	*RS2* (w/o group descriptions)	23	8
Others		7	5

[object function *pp/comp*] [object *np*] ga(SUBJ) [position *pp/comp*] [object description *np*] desu(COPULA).

Both patterns, which assert the features of the object including its position, handle the availability of the object as old information. Examples of *RS3* are

(14) onryou wo chousetsusuru botan ga
volume OBJ control button SUBJ
daiarubotan no hidariue ni arimasu.
dial-buttons upper-left LOC exist
'The button for controlling the volume is located to the upper left of the dial buttons.'

(15) sonotame no botan ga daiarubotan no
for-it button SUBJ dial-buttons
hidariue ni aru chiisai botan desu.
upper-left LOC exist small button COPULA
'The button for it is the small button to the upper left of the dial buttons.'

These patterns are used when the existence of a specific function or an object used for such a function was previously asserted. In those cases, as such an information is old, *RS3* is appropriate, while all other schemata described above are not. Although it must be possible to classify pattern *RS3* into smaller classes and to discuss the occurrence frequency and the situations in which they occur, the small numbers involved prevented further investigation.

2.5 Relation to previous research

The occurrence frequency of each schemata listed above supports the findings of our previous research summarized as **P1** and **P2** in the introduction. In *RS1* and *RS2*, which are basis of all schemata, the object position is conveyed almost always under the guidance of the schemata themselves. In particular, it is mandatory in *RS1*. So, the amount of information conveyed for identifying objects, how much is needed depends as a matter of course on the modes available, is controlled by feature descriptions other than position information. This causes **P1**, the property that the usage of position information does not decrease in MMD, while other kinds of information do decrease. In addition, this property is seen more strongly in MMD; *RS1"* and *RS2"* are used frequently wherein a group member directly is referred to directly to the object; the group is not mentioned.

In SMD, *RS1?* and *RS2?* are used more frequently than in MMD. This means that references to the surface where the object is located and the group it belongs to tend to be made in an utterance different from the utterance referring to the object itself. In addition, *RS*1* also appears more frequently in SMD than in MMD. This means an identification request and an action request are made separately. These are indications of **P2**, the property that actions tend to be realized as an explicit goal and divided into a series of fine-grained steps in SMD.

3 Actions coordinated with reference expressions

In MMD, several kinds of physical actions accompany referring expressions. Proper coordination between such physical actions and linguistic expressions makes the referring actions fluent. In addition, referring expressions in MMD frequently use demonstratives such as "kore(*this*)" and "koko(*here*)" in relation to these actions. Investigating the constraints or patterns of this coordination and applying them to the schemata of referring expressions makes it possible to generate fluent action statements.

Physical actions in referring actions in MMD are divided into the following three categories.

Exhibit actions: Actions for making object visible such as turning a body or opening a cassette cover[7].

[7] Exhibit actions contain both general actions like turning the body and machine specific actions like opening the cassette cover. There may be some differences between these two types of actions. For example, in referring expressions, the latter is usually requested directly by an imperative sentence, while the former is requested indirectly by directing attention to a specific side or implicitly by mentioning that side.

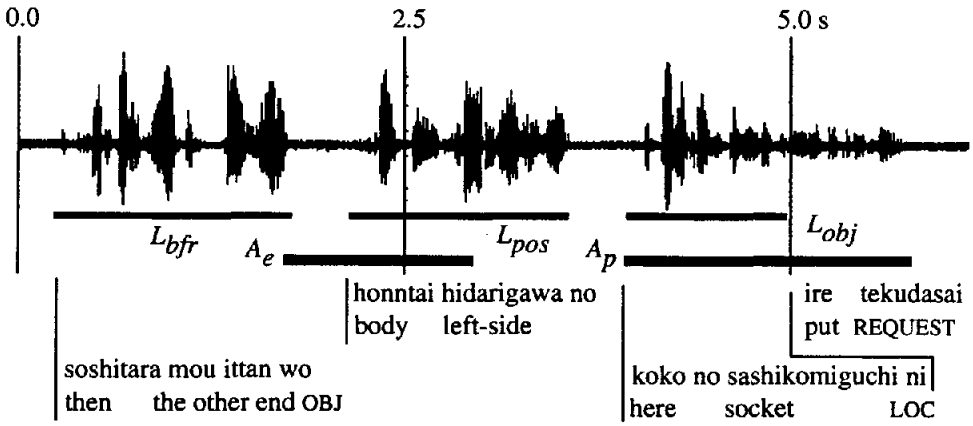

Figure 2: An example of temporal coordination of exhibit actions

Pointing gestures: Deictic actions pointing to/at objects

Simulated operations: Actions that are parts of machine operations such as pushing a button and picking up a handset. In instruction dialogues, experts sometimes just simulate these actions without actual execution.

This section reports the patterns of temporal coordination of these actions with linguistic expressions, based on the observation of the corpus. Videotapes of just 48 referring actions (4 experts referred to 12 objects once each) were examined. As the amount of data is so small, we provide only a qualitative discussion.

3.1 Exhibit actions

Only invisible objects need exhibit actions when they are referred to. Being those objects referred to, whichever scheme listed above is used, the information of the position or surface where the object is located is conveyed ahead of other features of the object. That is, letting the linguistic expression just before the referring expression be L_{bfr}, the position description be L_{pos}, and the object description be L_{obj}, the temporal relation of those can be summarized as follows using Allen's temporal logic (Allen, 1984).

$$L_{bfr}\ before\ L_{pos}\ before\ L_{obj}$$

Accompanying these expressions, exhibit action A_e, and pointing gesture A_p, have the following relations.

$$L_{obj}\ starts\ A_p$$

$$L_{bfr}\ before\ A_e\ before\ L_{obj}$$

$$L_{pos}\ overlaps \mid overlaps^{-1} \mid during \mid during^{-1}\ Ae$$

The pointing gesture to the object begins at the same time of the beginning of the object description. The exhibit action is done between the end of the utterance just before the referring action and the beginning of the object description. The exhibit action and position description relates loosely. There may be a temporal overlap between them or one may be done during the other. More precise relation than this could not concluded. In order to keep these relations, pauses of a proper length are put before and/or after the position description if needed.

Fig. 2 shows a schematic depiction of the above relations and an example taken from the corpus.

3.2 Pointing gestures and simulated operations

Pointing gestures are faithfully synchronized to linguistic expressions. During action statements, almost all descriptions of objects or positions are accompanied by pointing gestures. Landmarks and object groups are also pointed to. When a pointing gesture is not made to the currently described object, no pointing gesture is made. Pointing gestures to objects other than the currently described one never happen. One exception to this constraint is scheme *RS3*. When the subject part of *RS3*, which is an object description, is uttered, no pointing gesture is provided. A pointing gesture begins as the position description begins.

The linguistic description of an object, L_{obj}, and a pointing gesture to it, A_p, basically satisfy the temporal relation, L_{obj} *starts* A_p. That is, L_{obj} and A_p begin at the same time, but A_p lasts longer. However, the constraint mentioned above, that pointing gesture to objects other than currently described one never happen overrides this relation. As a result, in general, the pointing gesture to an object begins after finishing reference to other objects. As other

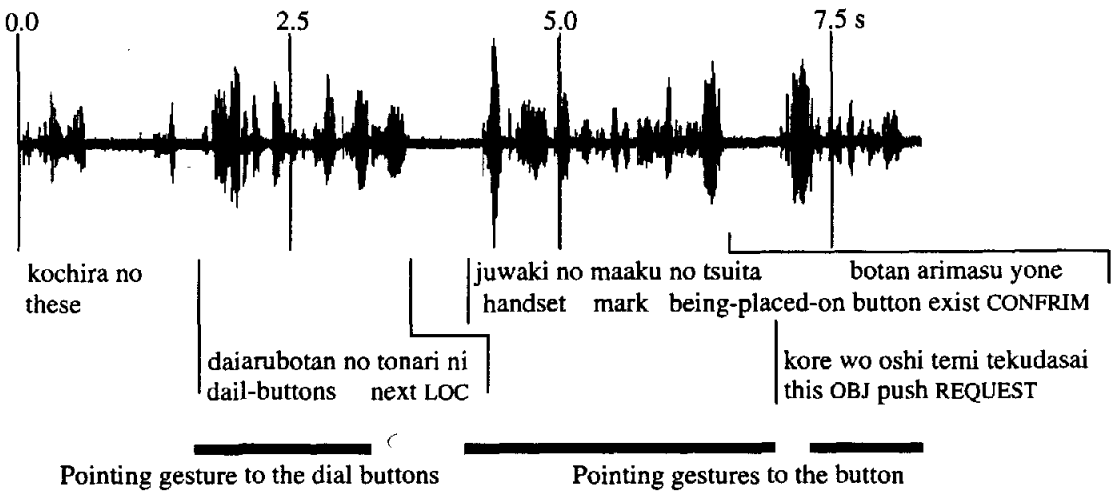

Figure 3: An example of temporal coordination of pointing gestures

objects are usually mentioned as landmarks for describing the object position, a pointing gesture to the object begins midway through position description. A_p usually lasts after L_{obj}. In particular, a pointing gesture to the main object of a referring expression lasts till the utterance ends and the addressee acknowledges it. So, in the case of *RS1*, a pointing gesture lasts till the end of the sentence that asserts object existence.

When more than one noun phrase or postpositional phrase describing the same object are uttered successively as in cases of appositive expressions, the pointing gestures are once suspended at the end of a phrase and resumed at the beginning of the next phrase. This is prominent when the next phrase begins with a demonstrative such as "this".

Simulated operations are synchronized with the verb describing the corresponding operation. Their synchronization is more precise than the case of exhibit actions. As a simulated operation such as button pushing is similar to a pointing gesture, a suspension and resumption similar to one mentioned above is done probably to distinguish them.

Fig. 3 shows an example taken from the corpus. In this example, it is not clear whether the last action is a pointing gesture or a simulated operation.

4 Discussion on implementation

We have began to implement a referring action generation mechanism using the schemata derived and coordination patterns described so far. The experimental system we are now developing shows a GIF picture of the telephone, and draws a caricatured agent over it. The pointing gestures are realized by redrawing the agent. As every picture is annotated by the object positions it contains, generating a pointing gesture and interpreting user's one are possible and easy. Other actions such as turning the body and opening the cassette cover are realized by playing a corresponding digital movie at exactly the same place as the GIF picture displayed[8]. The first frame of the digital move is the same as the GIF picture shown at that point of the time, and while the movie is being played, the picture in the background is switched to the one equaling the last frame of the movie. Fig. 4 depicts this mechanism. Those actions need considerable time as do human experts. This is in contrast to our previous system which implemented such actions by switching pictures so the time taken was negligible.

The framework adopted for coordination between utterances and actions is synchronization by reference points (Blakowski, 1992). The beginning and end points of intonational phrases must be eligible as reference points. It's still under consideration if just waiting, for which pauses are put after the action finished earlier, is enough or more sophisticated operations such as acceleration and deceleration i.e. changing utterance speed, are needed. The need for dynamic planning such as used in PPP (André & Rist, 1996) should be examined.

5 Summary and future work

What patterns of linguistic expressions are commonly used and how physical actions are temporally coordinated to them were reported based on cor-

[8]Tcl/tk is used for showing the GIF pictures and drawing the agent as well as other basic input/output functions; xanim is used for playing the digital movies.

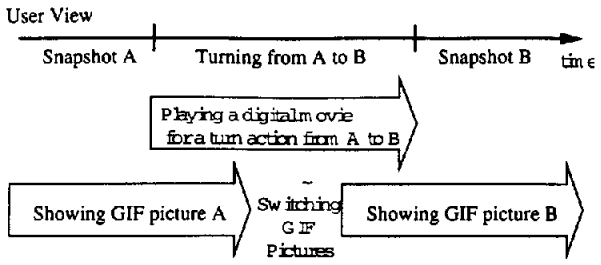

Figure 4: The implementation of turn actions

pus examinations. In particular, by categorizing objects according to two features, visibility and membership, the schemata of referring expressions could be derived. This categorization is still not sufficient for uniquely determining each reference expression, and some other features must impact the expressions used. This is, however, a good first step, as the two most dominant features were obtained. Moreover, the difference between the occurrence frequencies of those schemata in MMD and SMD explains the findings of our previous research. Implementation based on these results is on going.

There is a lot of future work beyond the implementation issues. First, the reported coordination patterns between linguistic expressions and actions must be verified in a quantitative manner. An objective criterion for annotating visual information is hard to establish. Overcoming this problem is important and unavoidable. Next, our research must be generalized in two perspectives: the results must be confirmed in many materials other than our telephone; the degree of dependency on the language used must be examined.

One of major problems stemming from our approach is that the importance of criteria is not clear. Although the criteria can be derived by observing and modeling the explanations made by experts, there may be fluent explanation techniques not yet observed. Deviations from the criteria do not cause a big problem, and the recipients do not perceive them to be awkward. These problems can be examined when the system currently implemented is made to generate several types of referring actions experimentally.

References

Allen, J.F., "Towards a General Theory of Action and Time", Artificial Intelligence, Vol. 23, No. 2, 1984, pp. 123 - 154

André, E. and Rist, T., "Coping with Temporal Constraints in Multimedia Presentation Planning", Procs. of AAAI-96, Vol. 1, pp. 142 - 147, 1996

Appelt, D.E., "Planning English Referring Expressions", Artificial Intelligence 26, 1985, pp. 1 - 33

Blakowski, G., Hüel, J., Langrehr, U. and Mülhäser, J., "Tool Support for the Synchronization and Presentation of Distributed Multimedia", Computer Communication, Vol. 15, No. 10, 1992, pp. 611 - 618

Cassell, J., Pelachaud, C., Badler, N. and et al., "Animated Conversation: Rule-based Generation of Facial Expression, Gesture & Spoken Intonation for Multiple Conversational Agents", SIGGRAPH 94, pp. 413 - 420, 1994

Cohen, P.R., "The Pragmatics of Referring and the Modality of Communication", Computational Linguistics, Vol. 10, No. 2, 1984, pp. 97 - 146

Clark, H.H. and Wilkes-Gibbs, D. "Referring as a Collaborative Process", "Intentions in Communication" eds. Cohen, P.R., Morgan, J. and Pollack, M.E., The MIT Press, 1990, pp. 463 - 493

Dalal, M., Feiner, S., McKeown, K. and et al., "Nagotiation for Automated Generation of Temporal Multimedia Presentations", Proc. of ACM Multimedia 96, pp. 55 - 64, 1996

Dale, R. and Haddock, N., "Content Determination in the Generation of Referring Expressions", Computational Intelligence, Vol. 7, No. 4, 1991. pp. 252 - 265

Feiner, S.K., Litman, D.J., McKeown, K.R, and Passonneau, R.J., "Towards Coordinated Temporal Multimedia Presentations", "Intelligent Multimedia Interfaces" eds. Maybury, M.T., The AAAI Press/The MIT Press, 1993, pp. 117 - 138

Kato, T. and Nakano, Y.I., "Referent Identification Requests in Multi-Modal Dialogs", Procs. of Int. Conf. on Cooperative Multimodal Communication, Vol. 2, pp. 175 - 191, 1995

Kato, T., Nakano, Y.I., Nakajima, H. and Hasegawa, T., "Interactive Multimodal Explanations and their Temporal Coordination", Procs. of ECAI-96, pp. 261 - 265, 1996

Kuno, S., "Danwa no Bunpou", Taishuukan Shoten, 1978, In Japanese