ANALYSIS AND INTEGRATION OF MULTIMODAL INPUTS IN INTERPRETING TELECOMMUNICATIONS

Kyung-ho Loken-Kim Suguru Mizunashi Mutsuko Tomokiyo
Laurel Fais Tsuyoshi Morimoto
e-mail kyungho@itl.atr.co.jp
ATR Interpreting Telecommunications Research Laboratories
2-2 Hikaridai Seika-cho Soraku-gun Kyoto 619-02
Japan

The study reported here is an attempt to understand human verbal-gestural behavior in a multimodal bilingual setting. Specific questions addressed are: 1) What kind of deictic gestures people use in a machine-mediated condition, and how these differ from those used in a human-mediated condition, 2) How significant the use of gestures is in each condition, and 3) How verbal and gestural behaviors are interrelated, 4) What the implications of our findings are for a multimodal spoken language interpretation system. In this paper, we attempt to answer these questions, and introduce the architecture of a prototype multimodal user interface system. This system takes spoken language and deictic gestures and produces a semantic representation of the inputs.

1. INTRODUCTION

For the past nine years, ATR's Interpreting Telecommunications Research Laboratories has been conducting research on the issues involved in enabling two people speaking different languages to communicate through an automatic interpretation system [1]. ATR's first speech-to-speech interpretation prototype, ASURA, was successfully demonstrated with good media reports. Human-to-human communications, nevertheless, rarely rely on the auditory channel alone; visual, and tactile interactions are all inherent elements of human communications. With this in mind, researchers at ATR have been exploring the possibility of introducing a new dimension - multimodality - to the spoken language interpretation system.

In a multimodal system, because visual objects are present, users have the option of incorporate them into the communication through some form of deictic gesturing [2]. In person to person speech, deictic gestures eliminate the need for a lengthy definite description and simplify the dialogues. It has been, therefore hypothesized that they will have a favorable influence upon spoken language interpreting systems because they will reduce the speech recognition workload. Gestures are, however, in many cases, ambiguous, incomplete, and sometimes impossible to understand without verbal and contextual information. In our previous research, we also found that speech when uttered in parallel with deictic gestures, often tends to break into fragments, and is, in many cases, incomprehensible without the information provided in the gestures. Therefore, intelligent mapping of the demonstratum (the region to which the user points) onto a referent (the region to which the user intends to refer), and a referent onto a descriptor (the descriptive part of the accompanying noun phrase or deixis) becomes an important issue for a multimodal spoken language interpretation system[*1]

The study reported here is an attempt to understand human verbal-gestural behavior in a multimodal bilingual setting. Specific questions addressed are: 1) What kind of deictic gestures people use in a machine-mediated condition, and how these differ from those used in a human-mediated condition, 2) How significant the use of gestures is in each condition, and 3) How verbal and gestural behaviors are interrelated, 4) What the implications of our findings are for a multimodal spoken language interpretation system.

In this paper, we attempt to answer these questions, and introduce the architecture of a prototype multimodal user interface system. This system takes spoken language and deictic gestures and produces a semantic representation of the inputs.

2. ANALYSIS OF MULTIMODAL INPUTS

2.1. Experimental Setting and Subjects

To answer these questions, we have conducted two experiments: one in a human-mediated (HM) setting, and one in a machine-mediated setting (Wizard of Oz (WoZ) method).

A total of 39 subjects (18 Japanese: 15 acting as conference agents, one as a "wizard" interpreter, and two as human interpreters; and 21 North American native speakers of English: 20 acting as clients, and one as a "wizard" interpreter) took part in the experiment. Clients were told to imagine that they had arrived at Kyoto Station for the first time and were trying to get information from the agent in

---

[*1] In this study, we only concern ourselves with one-to-one mapping of a demonstratum onto a descriptor and we assume the demonstratum-referent mapping is always correct. Definitions are from [3].

order to find their way to a conference.

Each subject (including agent, client, and interpreters) was provided with a computer display equipped with a touch panel. Subjects were allowed to write or mark on the map of Kyoto Station while verbally interacting with each other.

The client's and agent's utterances were interpreted by 1) human interpreters in HM, and 2) two "*wizards*" acting like a "*machine*" interpretation system in WoZ. One native speaker of Japanese acted as a "*wizard*", translating the English into Japanese, while another native speaker of North American English translating the Japanese into English. The "*wizards*" modulated their speech to be as monotonic and syllable-timed as possible, simulating a machine-generated voice. "*Wizards*'" voices were, also, distorted by a voice effector to make the subjects believe they were actually interacting with an interpreting machine rather than human interpreters.

During the training sessions, it became clear that the "*wizards*" were having difficulties in generating interpreted messages in a machine like tone, while regenerating gestures in a machine like manner. In order to lessen the "*wizards*'" burden, their task was simplified as follows. Client's (or agent's) gestures were, first, transmitted only to "*wizards*" (into a buffer), then, "*wizards*" choose an appropriate time to transmit the client's gestures as they interpret the verbal messages. The "*gesture transmit button*" on the screen allowed the interpreters to select an appropriate time to transmit the gestures. An experimenter monitoring the conversations instructed the "*wizards*" to ask the subjects to repeat an utterance during the course of the experiment when it was especially long, disfluent, or complex. The utterances by the "*wizards*", called "*repetition requests*" (RR), were usually took the forms "*Please repeat*" and "*Please speak slowly.*" Gestures which appeared on the screens of the interpreters (under the current system configuration, all gestures appeared on the interpreters' screens) were video taped and later analyzed.

3. RESULTS

3.1. Gesture Classifications

By analyzing the video tapes, we were able to classify gestures into four types: circling, line-dragging, pointing, and others.

A "circling" gesture typically encircles objects on the screen. In both HM and WoZ, agents used many more circling gestures than clients, and twice as many circlings were used in WoZ as in HM.

A "line-dragging" is a gesture which creates straight or curved lines on the screen. Line -draggings were used in four different ways: 1) to connect two points on the map with a sentence such as, <*Go from here to here*>*d*[*1], 2) to refer to a specific object by starting or ending the gesture near the object with a sentence such as, <*Please go all the way to here*>*d*, 3) to refer to an object by drawing a line near or over the object with a sentence, such as, <*Go out from this side*>*d*, and 4) to show a process, i.e., how to get from point A to point B, by a trajectory of the line with a sentence such as, <*You could go like this*>*d*." Like circlings, in both HM and WoZ, the agents used more line-draggings than the client, and the subjects used twice as many draggings in WoZ than in HM.

*1 Our transcription convention for deitic gestures: <underline>c:circling, <underline>d:line-dragging, <underline>p: pointing, <underline>m:marking

Table 1. Gestures during Human-Mediated Experiments

	circling	dragging	pointing	others	total
agent	18	25	0	2	45
client	4	4	3	1	12
total	22	29	3	3	57

Table 2. Gestures during Machine-Mediated (WoZ) Experiments

	circling	dragging	pointing	others	total
agent	42	48	0	1	91
client	3	13	11	8	35
total	45	61	11	9	126

The result show (Table 1, and 2) that circlings and line-draggings were used 86% of the time, but line-draggings were used more than circlings. This may be caused by the directional characteristic of the task. The rest of the gestures were used mainly for referent identifications.

3.2. Gestures and Turns

Table 3 and 4 are the summaries of the number of dialogue turns taken to accomplish the task in both conditions. Although it took fewer turns in WoZ (384 turns) than in HM (595 turns), the proportion of gestures to turns was greater by a factor of three in WoZ (32.8%) than in HM (9.5%) (Table 5). This suggests that there is a clear tendency for the subjects to rely on the visual channel whenever they faced communication difficulties (mainly caused by RR messages). This was evident in spite of the fact that they were not completely comfortable with the multimodal terminals. The results of both section 3.1 and this section show that the agents, as information providers, were much more active in using gestures than the clients, as information receivers. This concurs with our previous findings [4].

3.3. Verbal-Gestural Behavior

In this section, we describe the semantic, and temporal interdependencies between verbal descriptors (deixis, proper nouns) and gestures.

3.3.1. Descriptor-Gesture-Demonstratum Relations

We found that circling gestures always had corresponding verbal descriptors in the form of deixis, proper nouns, and adverbs. Out of 61 circling gestures, 41 circlings were accompanied with deixis, such as, "here" (ここ), "there" (そちら), "this side" (こちらの方), and the remaining 20 with proper nouns and adverbs (Table 6).

Line-dragging gestures, while most of them also had corresponding verbal descriptors in the form of deixis, proper nouns, and adverbs, sometimes did not have specific verbal descriptors. Out of 97 line-dragging gestures, 38 were accompanied by deixis, 59 by proper nouns and adverbs, but 4 were without descriptors.

Table 3. No. of Turns in Human-Mediated Condition

	agent	client	total
A	37	10	47
B	260	288	548
total	297	298	595

Table 4. No. of Turns in Machine-Mediated Condition

	agent	client	total
A	46	20	66
B	151	167	318
total	197	187	384

A: No. of Turns with gestures,
B: No. of turns without gestures

Table 5. Percentage of Gestures in Turns

	agent	client	overall
HM	15.1%	4.0%	9.5%
MM	46.2%	18.7%	32.8%

3.3.2. Descriptor-Gesture-Demonstratum-Temporal Interdependencies

We further investigated the temporal interdependencies among descriptors, gestures, and demonstratums. As was mentioned, circlings always had corresponding verbal descriptors. 45% of the circling onsets coincided with the onsets of their verbal descriptors (1 in Figure1). On the other hand, 41% of the circling offsets coincide with the onsets of the descriptors (2 in Figure 1).

For line-draggings, in 40.9% of the gestures, the subjects drew lines on the object (3 in Figure 2), and they started well before the onsets of the verbal descriptors. On the other hand, 30% of the gestures ended at the onsets of the verbal descriptors, and 22.7% of the gestures were drawn throughout the verbal descriptor for the purpose of describing "process."

4. INTEGRATION OF VERBAL-GESTURAL INPUTS

In addition to the goal of understanding human verbal-gestural behaviors, this study aimed at developing a multimodal human-computer-human interface that could be added to ATR's current speech-to-speech interpretation system. In this section, we briefly describe the six modules

Table 6: Gestures and Demonstratums

Total number of circling gestures: 61
  Circlings with deixis: 41
    demonstratum: location: 39
    demonstratum: object: 2
  Circlings with other than deixis: 20
    demonstratum: location: 20
Total number of line-dragging gestures: 97
  Line-dragging with deixis: 38
    demonstratum: location, object near the onset of the gesture: 8
    demonstratum: location, object near the offset of the gesture: 11
    demonstratum: location, object on the path of the gesture: 14
    demonstratum: process (e.g., how to get from A to B): 5
  Line-dragging without deixis: 59
    demonstratum: location, object near the onset of the gesture: 3
    demonstratum: location, object near the offset of the gesture: 19
    demonstratum: location, object on the path of the gesture: 22
    demonstratum: process: 15
  Line-dragging without descriptor: 4

Figure 1. Circling and Verbal Descriptor

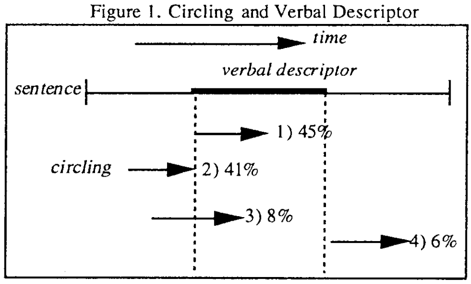
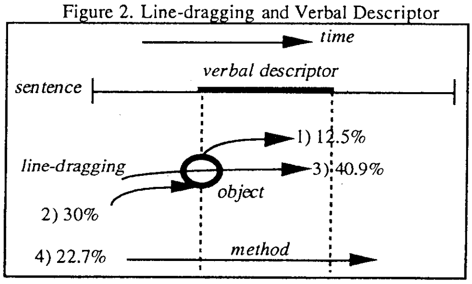
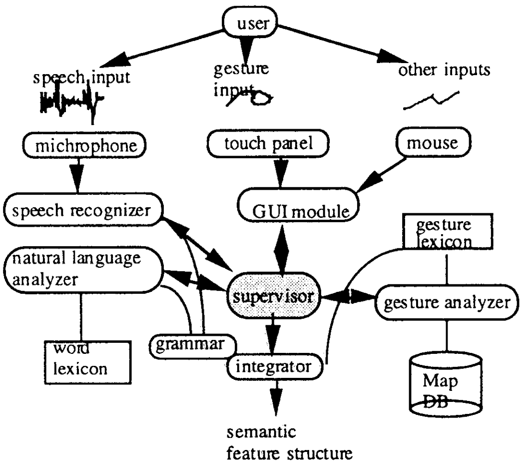


Figure 2. Line-dragging and Verbal Descriptor



of our first prototype multimodal interface system (Figure 3). This system takes multimodal inputs, integrates them in a temporally well coordinated manner, and produces a unified semantic representation of the inputs.

Figure 3. System Overview



(A) Word Lexicon, Gesture Lexicon, Map Database, Grammar

a) Word Lexicon

Currently there are 43 words related to the direction finding tasks in the word lexicon. Words and their attributes are represented in a feature structure, and deixis are augmented with temporal information to capture the acoustics, and spatial information the gestures (Figure 4).

Figure 4. Deictic Expression Feature Structure

```
(deflex-named このあたり-1 このあたり n-deictic
!(lex-phon-orth "konoatari" "このあたり")
(<!m sem> == [[RELN DEITIC-PLACE]
              [AGEN *SPEAKER*]
             [RECP *HEARER*]
              [OBJE [[RELN このあたり]]]])
(<!m time-stamp> == [[SPEECH [[tS ?X1]
                               [tE ?X2]]]])
(<!m gesture> == [[RELN CIRCLING-3]
                  [LOCATION [[IS [[X ?X][Y ?Y]]]
                  [TIME-STAMP
                  [[mouse [[tS ?X1][tE ?X2]]]]]])
(!m !prag> == [[iterr agen]]))
```

b) Gesture Lexicon

Currently, there are only eight entries in the gesture lexicon. Each entry is also a feature structure of a gesture (Figure 5) ranging from circling to line-dragging. Features in the structure are designed to capture the temporal-spatial information of the gestures.

Figure 5. Gesture Feature Structure

```
(deflex-named CIRCLING-3 CIRCLING gesture
  (<!m sem> == [[RELN CIRCLING-3]
     [LOCATION [[ lS [[X ?X][Y ?Y]]]
                 [lE [[X ?X][Y ?Y]]]]]
     [TIME-STAMP [[mouse [[tS ?X1][tE ?X2]]]]]])
```

c) Map Database

Objects on the map are represented with a list of attributes as follow (Figure 6).

Figure 6. Representation of the Map

```
[Object number][min X][min Y][max X][max Y]
     [kind of object][name of object]

example: [1][56][145][70][178][hotel][kyoto-hotel]
```

d) Grammar

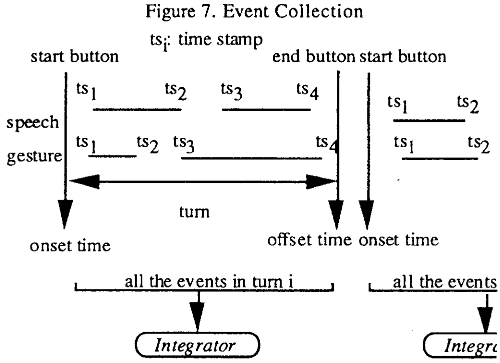
Currently, there are 114 grammar rules used both for the speech recognition and language analysis. The vocabulary size is 43 words, with a phoneme perplexity of 1.74.

(B) Supervisor

Supervisor is a multifunction module that controls all the sub-modules and regulates data flow. When the user starts a "*turn*" by selecting the "***start /stop button***", the *supervisor* 1) initializes all the modules (at this point, the speech recognizer starts taking speech input), 2) reads the system clock time ("*onset time* "*(start time)*) and notifies it to each module. During a "*turn*", the *supervisor* 3) receives data from one module and transmits it to another module. For example, each time the user completes a gesture, all the x, y coordinates are read from the *gesture analyzer* and they are transmitted to the *integrator.* When the user ends a "*turn*" by selecting the "*start/****stop*** *button*", the *supervisor, then* 4) notifies the ending of a "*turn*" to all the modules (the speech recognizer stops taking speech input and generates recognition results). Users can terminate the entire process by selecting the "*quit button.* " Upon receiving the termination command, *supervisor* 5) deactivates all the modules, and 6) releases all the relevant resources (memory, temporary files etc.).

One of the most important functions of the supervisor, however, is the "*event collection*", that is, collecting all the peripheral events (speech, gesture, etc.) that took place in one turn (between "*onset time*" and "*offset time*") and handing them over to the *integrator* (Figure 7).

Figure 7. Event Collection



(C) Speech Recognizer

We have adopted a continuous speech recognizer which is based on a phone-synchronous SSS-LR [5] technique developed at the ATR Interpreting Telecommunications Research Laboratories. This speech recognizer was developed with emphasis on modularity so that new modules could easily be added. The recognition accuracy varies from 85% to 92% depending on the number of states in HMnet.

Sentences recognized are mostly short and simple, and they contain instances of deictic expressions, such as, 京都ホテルは<u>このあたり</u>です (Kyoto hotel is <u>around here</u>). Sentences can be uttered in either continuous or connected mode; users are free to utter a sentence in one breath, or leave a pause between two bunsetsu phrases. The output from the recognizer is a triplet: recognized word, onset time, and offset time for each word (Figure 8).

Figure 8: Speech Recognition Results

```
                    sentence:
              京都ホテルはこのあたりですか

                recognition results:
1135          : time elapsed since the turn "onset time
京都ホテル      0    830 (speech onset & offset time)
は              830  920
3842          :time elapsed since the turn "onset time"
このあたり      0    780: speech onset time reset
                          due to the pause
で              780  860
す              860  1050
か              1050 1200
京都ホテルはこのあたりですか -32.115994
```

(D) Natural Language Analyzer

The *natural language analyzer* was developed using a parsing toolkit [6]. This parsing toolkit was developed with emphasis on efficient unification and modularity to handle many of the linguistic phoneme in spontaneous speech. The input to this module is the results of the speech recognition. Upon receiving the recognition result, the *natural language analyzer* first generates a parse tree using the grammar rule, then converts the tree to a dependency structure, and finally produces a semantic feature structure of the utterance (Figure 9). The feature structure is then handed over to the *integrator*.

(E) Gesture Analyzer

The main functions of the *gesture analyzer* are: 1) recognizing the kind of deictic gesture (circling, line-dragging, etc.,), 2) selecting demonstratums (objects[*1]), and 3) generating a temporal-spatial information of the gesture (Figure 10).

Algorithms for recognizing the kind of gesture, and identifying demonstratum are as follows [7]

<u>(1) Recognizing gestures</u>

a) Save entire trajectory points (x, y coordinates) of a gesture.

b) Compute the minimum and maximum values of the coordinates (Figure 11) and find its center point.

c) Divide the area into 8 regions, and divide the coordinates that belong to each reason.

d) If coordinates exist in every region, and the Euclidian distance between the onset and offset of the gesture is less than 50 (currently assigned value), then the gesture is a circling.

Figure 9. Output of the Natural language Analyzer

```
   sentence: 京都ホテルはこのあたりですか
[SEM [[RELN *YN-QUESTION*]
  [AGEN *SPEAKER*]
  [RECP*HEARER*]
  [OBJE [[RELN *BE-LOCATED*]
     [IDEN [[RELN * 京都ホテル* ]]]
     [PLACE [[RELN *DEICTIC-PLACE*]
        [AGEN *SPEAKER*]
        [RECP *SPEAKER*]
        [OBJE [[RELN * このあたり* ]
           [PRAG [[ITERR *SPEAKER*]]]
           [SYN [[POS NP]
              [INDEX [[EXTENT +]
                 [PARTIAL +]]]]]
           [TIME-STAMP [[SPEECH [[tS ?X1]
                        [-tE ?X2]]]]]
           [GESTURE ?GESTURE]]]]]]]]
```

Figure 10. Temporal-Spatial Information of a Gesture

```
3: turn I.D.
circle: gesture analysis result
3119: gesture onset time
4864: gesture offset time
(897,921) (128,164): object coordinates
(X1,Y1) (X2,Y2)
```

e) If there are points in only one region, then the gesture is a pointing.

f) If there are no points in the region 6 and 7, and the Euclidian distance between the onset and offset of the gesture is less than 3 (currently assigned value), then the gesture is a marking.

g) Rests are line-dragging gestures.

<u>(2) Object selection</u>

a) Circling gesture: among all the objects that are either within or on the perimeter of the circle, the one that is closest to the center is selected.

b) Pointing gesture: the object located at the demonstratum is selected.

c) Line-dragging: the object located on the trajectory is selected.

d) Marking: the object nearest to the center is selected.

(F) GUI (Graphic User Interface) Module

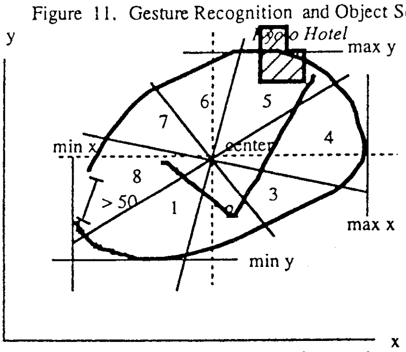
GUI module manages the user interface by displaying graphics (Figure 12), and monitors screen events (e.g., gestures on the touch panel) on the screen. Specifically, it 1) displays the map and other graphics, 2) reads the coordinates corresponding to the gesture trajectory on the map, 3) detects push-button events and 4) displays the result of temporal matching between the speech recognition and gesture, and 5) presents a unified semantic representation of the utterance and gesture.

(G) Integrator

The *integrator* 1) receives the semantic feature structure of

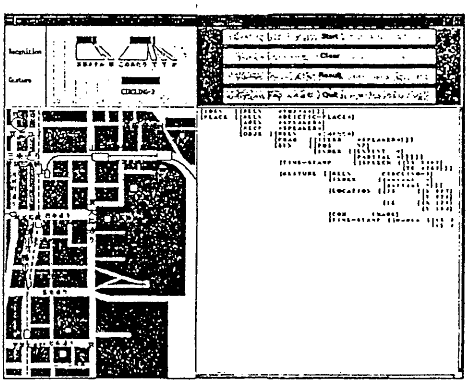
[*1] Currently, objects are not selected.

Figure 11. Gesture Recognition and Object Selection



numbers are region numbers

Figure 12. User Interface



the utterance from the *natural language analyzer and* temporal-spatial information of the gesture from the *gesture analyzer*, 2) searches for the deictic features in the feature structure, 3) checks temporal alignments between deixis and gestures, and 4) instantiates the deictic features with the temporal-spatial values of the gestures (Figure 13). Deixis-gesture alignment is done from the beginning of the utterances and gestures are assigned one gesture to one deixis, with any residual gestures being ignored.

5. DISCUSSIONS AND CONCLUSION

How do we incorporate these findings into the design of a multimodal spoken language interpretation system? Our artificial impediment of user-computer communication (by RR messages) may not have reflected the true picture of future man-machine communication, it, nevertheless, enabled us to get a glimpse of human verbal-gestural behaviors in a multimodal spoken language interpretation system. Through this study, we were able to classify gestures used in multimodal situations, and we observed a sharp reduction in the number of verbal interactions and a high frequency of modality shifts and deictic gesture deployment in our simulation of man-machine multimodal communications. The reduction of the number of verbal interactions, of course, means less burden on speech recognition and language translation, and is a blessing considering the current status of the spoken language interpretation technology. However, the increase of gestures that accompanies this verbal reduction presents us with a new challenge; the complexity of verbal-gestural mapping and synchronizing in a bilingual context. We know that automatic interpretation of truly spontaneous speech is a formidable task. We also know that multimedia technology is attractive and enhances human-to-human communications. Seamless fusion of the two technologies presents an even greater challenge but we are beginning to explore through our prototype of a multimodal interface.

Figure 13. Semantic Representation of the Utterance Generated by the Natural Language Processor

```
[SEM [[RELN *YN-QUESTION*]
 [AGEN *SPEAKER*]
 [RECP *HEARER*]
 [OBJE [[RELN *BE-LOCATED*]
     [IDEN [[RELN * 京都ホテル* ]]]
     [PLACE [[RELN *DEICTIC-PLACE*]
         [AGEN *SPEAKER*]
         [RECP *SPEAKER*]
         [OBJE [[RELN * このあたり* ]
             [PRAG [[ITERR *SPEAKER*]]]
             [SYN [[POS NP]
                 [INDEX [[EXTENT +]
                     [PARTIAL +]]]]]
             [TIME-STAMP [[SPEECH [[tS 3482]
                     [tE 4262]]]]]
             [GESTURE [[RELN CIRCLING-3]
                 [LOCATION [[IS [[X 897]
                         [Y 921]]]
                     [IE [[X 128]
                         [Y 164]]]]]
                 [TIME-STAMP
                     [[mouse [[tS 3119]
                             [tE 4864]]]]]]]]]]]]]]]]
```

6. REFERENCES

[1] Tsuyoshi Morimoto, Masami Suzuki, Toshiyuki Takezawa, Gen'ichiro Kikui, Masaaki Nagata, Mutsuko Tomokiyo, "A Spoken Language Translation System: SL-TRANS2," COLING'92, pp. 1048-1052, 1992

[2] Wolfgang Wahlster, "User and Discourse Models for Multimodal Communication," in Intelligent User Interfaces by Joseph W. Sullivan and Sherman W. Tyler (eds.) pp. 45-67, acm press

[3] H. Clark, R Schreuder, and S. Buttrick, "Common Ground and the Understanding of Demonstrative Reference," Journal of Verbal Learning and Verbal Behavior, 22, 245-258 in [2]

[4] Kyung-ho Loken-Kim, Laurel Fais, Tsuyoshi Morimoto, "Multimodal and Telephone-only Dialogues in English and Japanese," IPSJ, Spoken Language Information Processing Workshop 5-4, Feb. 1995

[5] Harald Singer, Tomohilo Beppu, Atsushi Nakamura, Yashinori Sagisaka, "A Modular Speech Recognition System Architecture," Proc. of Acoust. Soc. Japan, Fall, 1994

[6] Toshihisa Tashiro, Tsuyoshi Morimoto, "A Parsing Toolkit for Spoken Language Processing," WGNL Meeting of IPSJ, 95-NLP-106, 1995

[7] Young-duk Park, Kyung-ho Loken-Kim, Laurel Fais and Suguru Mizunashi, "Analysis of Gesture Behavior in a MM Interpreting Experiment," ATR Tech. Report TR-IT-0091, March, 1995