TR-IT-0172

EMMI Progress Report: An Evaluation of Research Done with the First EMMI Interface

Laurel Fais, Suguru Mizunashi,
and Kyung-ho Loken-Kim

June 1996

Three experiments (same language, human-human; bilingual, human-interpreted; and bilingual, machine-interpreted) were conducted in the context of the first EMMI interface. All of the analyses done on the data gathered in these experiments are collected and summarized here. The effectiveness of the multimedia component of EMMI is assessed in relation to telephone mode; it is also evaluated as an interface for automatic language processing. Conclusions drawn from these evaluations and from direct examination of the system itself form the basis for the design of a second interface, also described here.

Table of Contents

Introduction[1]

The ATR Environment for Multimodal Interaction (EMMI) was designed as a multimedia interface between users who do not speak the same language. We have been investigating the possibility of integrating EMMI with an automatic translation system. The intent behind such an integration is to exploit the capabilities of a multimedia (MM) system in order to lessen the burden on the language processing system. A series of three experiments was carried out in order to collect data which could yield insights into the maximally effective configuration of the two systems. These experiments involved human-human, same language communication; two-language human-interpreted communication; and two-language "machine"-interpreted communication (in a Wizard-of-Oz setting). Subjects in the experiments conversed via the MM system and via the telephone. Detailed descriptions of the experiments are available in (Fais, 1994b; Fais *et al.*, 1995; Park *et al.*, 1994; Park *et al.*, 1995); a detailed description of EMMI appears in the Appendix. A number of specific analyses were made using the data collected.

The intent of this report is to summarize the findings of these analyses in order to examine the effectiveness of the first version of EMMI. The most crucial question for such an examination is, of course, the criteria by which effectiveness is to be measured. We took two general approaches to establishing such criteria. First, we wanted to examine the effectiveness of the MM system *per se*, that is, how it compared to telephone communication in a general sense. Thus, we looked at how much information subjects exchanged in each communication environment, and at how many words they used to do it, which gave us an idea of the efficiency of each communication environment. Thinking that one other benefit of a MM system may be the willingness of subjects to use it, we also examined subjects' reactions to the use of the MM system via a post-interview questionnaire.

Our second approach was to examine the effectiveness of the MM system for a specific purpose, that is, reducing the burden on a language processing system. In order to do that, we measured disfluencies in the conversations in both the MM and the telephone environments. We also looked at the raw number of words used in each environment. Further, we looked at the amount of lexical accommodation occurring in each setting; greater levels of lexical accommodation could allow more accurate prediction of sound strings in a speech recognition system. This approach also implies comparison with the telephone setting, but in this case, the comparison is made with respect to the criterion of effectiveness vis-à-vis integration with an automatic language processing system.

Below, we present the results of several different analyses done with the data collected in the EMMI experiments. Each of these analyses is pertinent both to a general comparison of MM and telephone settings and to the integration of MM and machine translation. However, we have placed each under the category that seems the most relevant.

First, we look at the balance between the use of speech and gesture for conveying information in the MM setting, and at the amount of information conveyed in the MM and telephone settings. Second, we give the results of the post-experiment interviews conducted to ascertain subjects' attitudes about using a MM system. In these first two areas, the emphasis is on the general comparison of MM and telephone settings. Third, we summarize disfluency results for the three experiments and finally, we look at lexical accommodation. These last two topics focus on the possible effects of integrating a MM system with an automatic language processing system. Some of the results discussed are taken from proceedings papers for various conferences (with references noted); some are reported for the first time here.

---

[1]The authors would like to thank Kazuhiko Kurihara for his hard work in implementing the new EMMI system and for writing the Appendix for this report.

In conclusion, we present the results of a slightly different kind of "experiment." Five expert computer users were asked to negotiate the standard tasks in the EMMI environment. Their responses and suggestions were taped and collated, and, along with the results discussed above, formed the basis for the design of a second version of the EMMI interface. The rationale and design of this system will also be discussed.

The Use of Speech and Graphics to Convey Information[2]

Motivation

The balance between the use of speech and graphics is pertinent both to a comparison of MM and telephone settings and to the integration of MM with language processing. The analysis described below measures the number of words used in each setting. If the number of words is lower for the MM setting, we can say that integration of a MM component with a language processing system lessens the processing burden on the system. Further, the efficiency of each conversational setting is examined, taking into account the information exchanged in each experimental condition. This allows us to compare the general effectiveness of the multimedia and the telephone settings.

Background

The old adage about a picture being worth a thousand words captures two important concepts in the field of multimodal communication technology. The first is that certain modalities are more appropriate than others for conveying certain types of information. For example, if the location of a building is to be conveyed, it is generally thought that a visual image such as a map will convey that information better than a verbal description such as "three hundred meters west of the intersection of Elm and Vine, on the north side of the street." A single picture seems to convey at once what a fairly complex grammatical structure takes longer to express.

This first concept plays a major role in a number of systems that are concerned with the automatic construction of multimedia presentations (André and Rist, 1993; Arens *et al.,* 1993; McCaffery *et al.,* 1995). These systems contain rules which match features of the information to be conveyed with the corresponding capabilities of the various media available and select the most appropriate medium for conveying that particular information. These systems recognize that a number of media may be appropriate, but in most cases, they contain heuristics for choosing one medium.

The second concept embodied in the adage concerns this choice. "One picture *is worth* a thousand words" implies that the picture should *replace* the use of words. This is the assumption behind choosing to represent, say, the location of a restaurant on a map, *instead of* describing it in words. The adage implies that, where you can use a picture, you don't need the words.

It is natural, then, that when we turn from multimodal presentation generation to the *use* of multimodal systems, we assume that humans will operate in much the same way. We might predict that if users have the *option* of presenting the location of a building on a map, they will use that option *instead of* presenting the information in some other way, for example, by typing or speaking the information. We might suppose that users will employ pictures (where that is the appropriate medium) instead of words.

This view is attractive in the field of natural language processing. Fully automatic real-time language processing has proven to be too difficult a goal to pursue for the near future. For that reason, systems designers have turned toward the use of multimedia options as a way to supplement language processing systems. If users in fact do employ the non-speech options available in such integrated systems instead of using (only)

---

[2]This analysis appeared in Fais and Loken-Kim (1995).

speech, that would reduce the amount of language processing necessary. It may then be possible to build a language processing system capable of handling such a reduced load.

Hypotheses

The "one picture is worth a thousand words" approach is applicable to the integration of EMMI with real-time, automatic machine translation. In a telephone setting, users have only speech available for conveying information, but in the MM setting, it is possible to use additional media options. If users convey information in modes other than speech, the translation system's job becomes that much easier. That is, if locations and directions are presented visually, and basic personal information is typed in, less information is presented in speech form and the speech translation system will carry less of the burden of communication. This provides strong motivation for an integrated *multimedia* translation system, rather than a telephone based system.

Our initial hypothesis is, then, that the integration of non-speech media in a communication environment will reduce the amount of speech used, when compared to a speech-only (telephone) setting. Below we report on the results of three experiments conducted in EMMI to test this hypothesis. By comparing the communicative behavior of subjects across varied interpretation conditions, we were able to assess the contribution made by the use of speech and that made by the use of non-speech media to each condition. By comparing communication behavior in the telephone and multimedia settings, we were also able to determine whether, in fact, the use of non-speech media in some sense *replaces* speech in conveying information, and results in a reduced amount of speech in a multimedia setting.

In the analysis of the conversations, we made three measures. First we counted the number of words in each conversation. This gives us a basic comparison of the use of telephone and MM settings. Second, we observed that in the multimedia conversations, there was a noticeable amount of conversation concerned with the mode of presentation itself (see below). We identified and labeled this kind of conversation. Third, we examined transcripts for "information units." A task analysis of the direction-finding portion of the experiment was made and a list compiled of all the possible "information units" that could be conveyed. Examples of such "units of information" are: location of the client in Kyoto Station; location of the bus stop; length of bus ride; amount of train fare; name of train line, and so on. Each conversation was analyzed to discover how many of those units it contained.

Results

Before we turn to the actual results of the experiments, let us look at what our hypothesis implies the results should be. If the use of non-speech media such as drawing on a map or typing replaces speech, we should find coordinated changes in the amount of information conveyed by speech and that conveyed by visual gestures (i.e., drawing or typing) as illustrated in the hypothetical Figure 1.

If our hypothesis is correct, the amounts of information conveyed by gesture and by speech should be inversely proportional: as the number of gestures increases, the number of words should decrease.

However, we found this not to be the case. Instead, a comparable graph reflecting the actual results of the experiments is shown in Figure 2.

A short digression to explain the construction of this graph is in order. Clearly, Figure 2 implies some notion of the worth of gestures relative to words. Although it is meant as illustrative only, and is not meant to make a claim about the relative weight of gestures and speech, the weights were not assigned completely arbitrarily. Natural language descriptors for a number of the gestures found in our experiments were constructed and the average number of words per descriptor was determined. These were cases of what

we call "deictic" gestures in which the gesture was related to a deictic expression in the speech of the gesturer (see below). An example would be "I am standing *here*" accompanied by a mark. This gave us a rough idea of how many words a deictic gesture might replace; it turned out to be eight. However, only about half of the gestures observed were deictic; the rest did not replace speech but instead were redundant. Thus, each gesture was weighted as four words. However, note that this was done for the purpose of illustration only.[3]

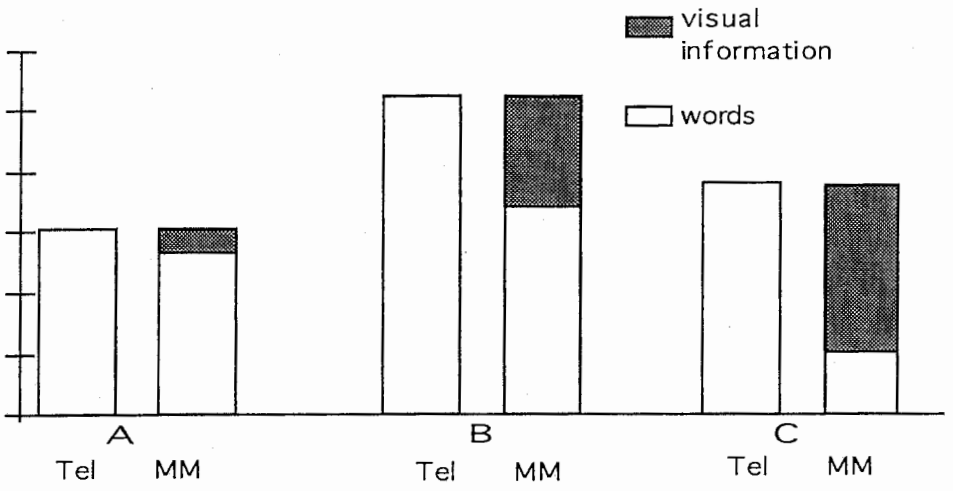

Figure 1. **Hypothetical** contributions of visual gestures and words in telephone and multimedia settings.

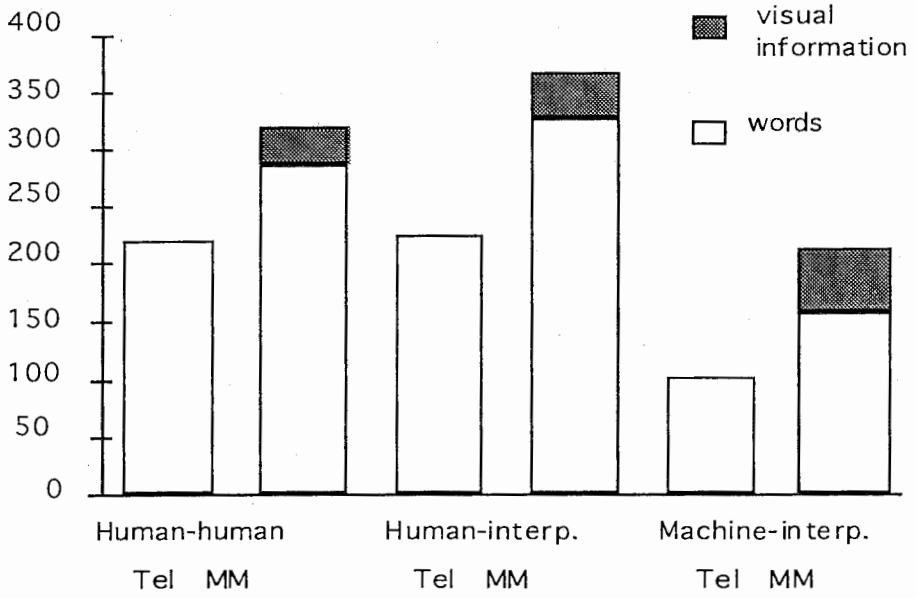

Figure 2. **Actual** contributions of visual gesture and words in the telephone and multimedia settings of the three experiments.

The actual weight assigned to the gestures used is irrelevant. Whatever the weighting system, the important point to note from this graph is this: in all three conditions, there was a significantly higher number of words in the *multimedia* setting than in the telephone setting. The use of gesture made a contribution to the information in the conversation above and beyond this greater number of words. This trend is opposite to what our hypothesis would suggest.

---

[3] Besides, "A picture is worth four words" just doesn't have the right ring.

Meta-media conversation. How can we explain these results? A closer examination of the conversations in the multimedia setting revealed examples like these:

1a. Agent: I'm circling the station...

1b. Client: I'd like to do whoop sorry whoop /laugh/ I'm sorry I remember something about you need to go up [uh] it's a little different cause that other one you can go up and /typing/ OK so and return

2a. Agent: and now we'll show you where you're goinu go

2b. Client: yes I was going to type in a message on the bottom

3a. Agent: and I can draw up a schematic of the bus station if you would like would you like to see the bus station

3b. Client: OK should I tell you also or just type it

4a. Agent: the Shinkansen is here I circle it for you can you see

4b. Client: can you see my location now I'm by the Shinkansen concourse

In (1), the speakers talk about what they are currently doing with the media; in (2), they talk about what they will do next; in (3), they ask their partners what they should do next; in (4), they confirm their partners' understanding of what they have just done. Each of these examples includes meta-conversation specifically concerned with managing the media available. We surmised that it was the addition of meta-conversation like this, what we call "meta-media" conversation, that was responsible for the unexpected increase in the number of words in the multimedia setting. Meta-media conversation is virtually absent from telephone conversations; however, there is a significant amount of this kind of conversation in the multimedia setting (Figure 3).

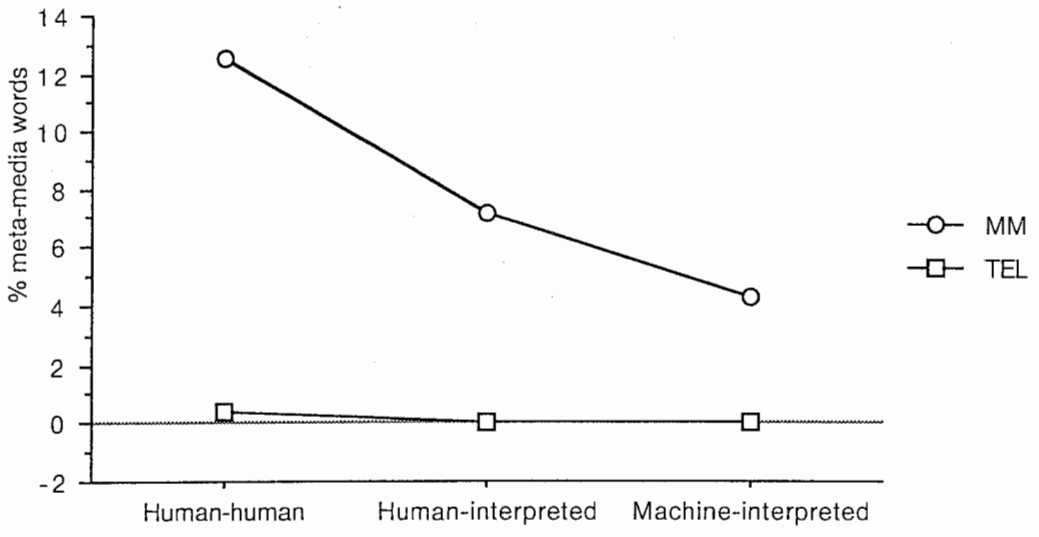

Figure 3. Percent of meta-media conversation in telephone and multimedia settings for all three experiments.

We then eliminated the meta-media conversation from our evaluation of the relative contributions of speech and visual gestures. However, even when meta-media conversation was subtracted out, the multimedia setting still showed a higher number of words than the telephone setting. This difference is no longer statistically significant for

the human-human experiment, but it is still significant for the human-interpreted and machine-interpreted conditions (Figure 4).

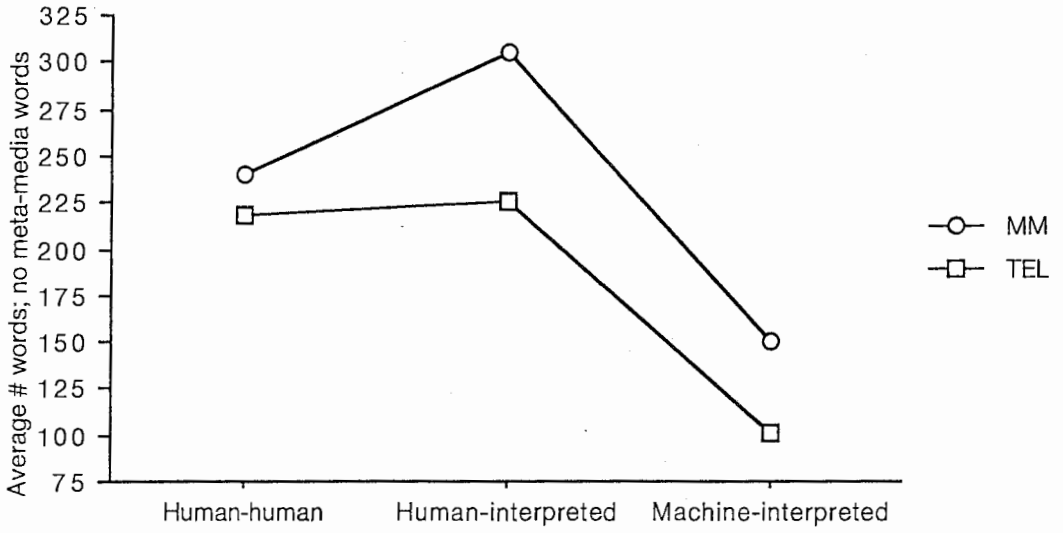

Figure 4. Number of words in telephone and multimedia settings for all three experiments, with meta-media words removed.

This suggests that, while meta-media conversation accounts for the "extra" words in the human-human condition, some additional factors are at work in the human-interpreted and machine-interpreted conditions.

Information units. What if clients are simply requesting and receiving more information in the multimedia setting of these conditions? This would have the effect of increasing the number of words used. In fact, there tends to be a higher number of information units in the multimedia setting for all three experiments (Figure 5), although this difference is not significant for any of the experiments. Because of this lack of significance, a greater amount of information cannot explain the higher number of words in the multimedia setting of the two interpreted experiments.

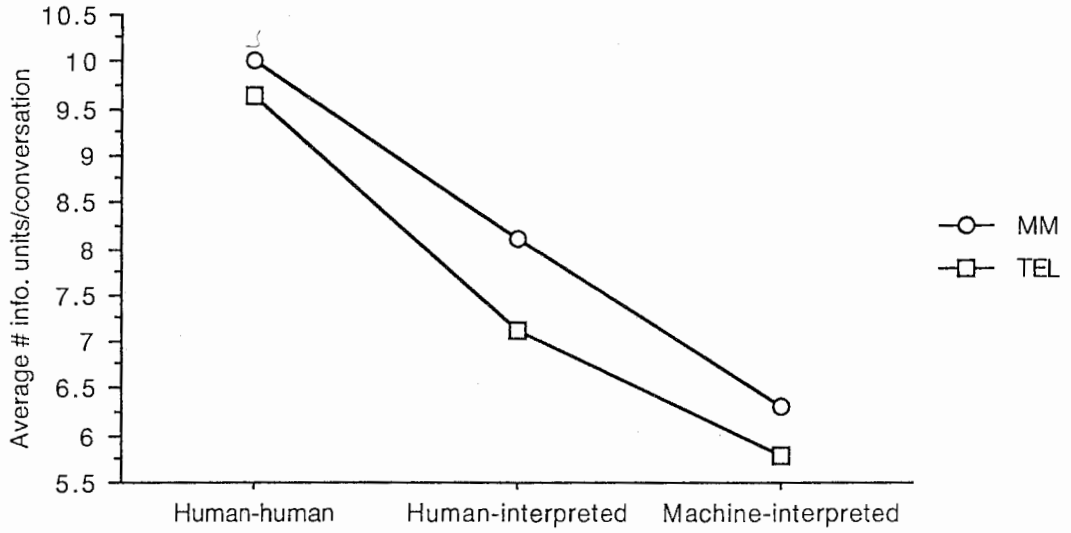

Figure 5. Number of information units in telephone and multimedia settings for all three experiments.

What might be more telling, however, is not the raw number of words used or of information units achieved, but the relationship between the two. Perhaps the number of words per information unit conforms to the expected trend (lower for multimedia; higher for telephone). When we examined the number of words used per information unit, however, we found the by-now familiar pattern: there is a significantly higher number of words used to achieve information units in the multimedia setting across all three experiments (Figure 6). We are faced with the same dilemma: in requesting and receiving information, subjects use more words in the multimedia setting (per information unit) than in the telephone setting.

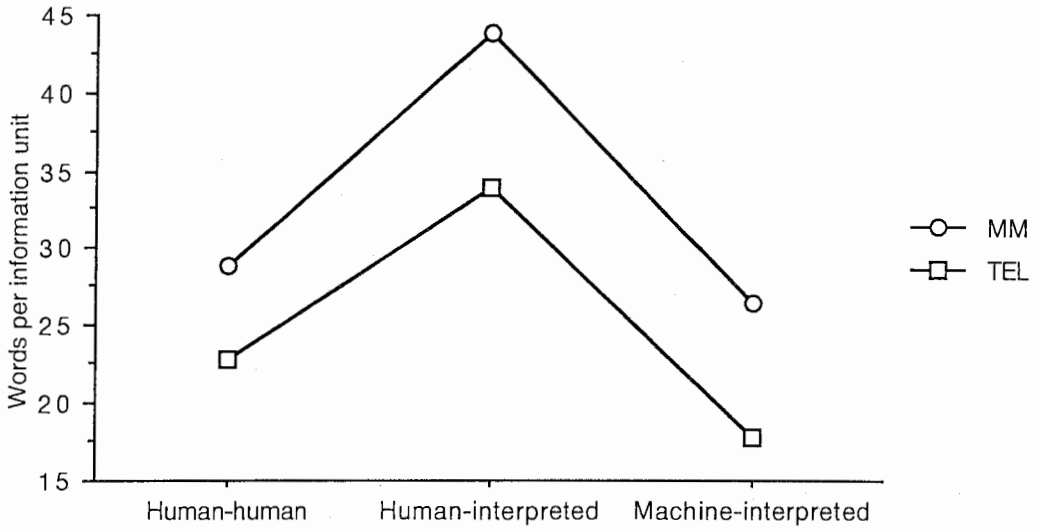

Figure 6. Words per information unit in telephone and multimedia settings for all three experiments.

Relationship of meta-media conversation and information units. We have examined the effects of meta-media conversation and the words-per-information-unit separately. What if we analyze the joint effect of these two factors on subjects' linguistic behavior in the telephone and multimedia settings? When we extract the meta-media conversation from the number of words, and then determine the number of words used per information unit, we find that the modal difference is no longer statistically significant. That is, if we ignore the meta-media conversation which takes place in the multimedia setting, the numbers of words used per information unit in both multimedia and telephone settings are equivalent in the two experiments (Figure 7).

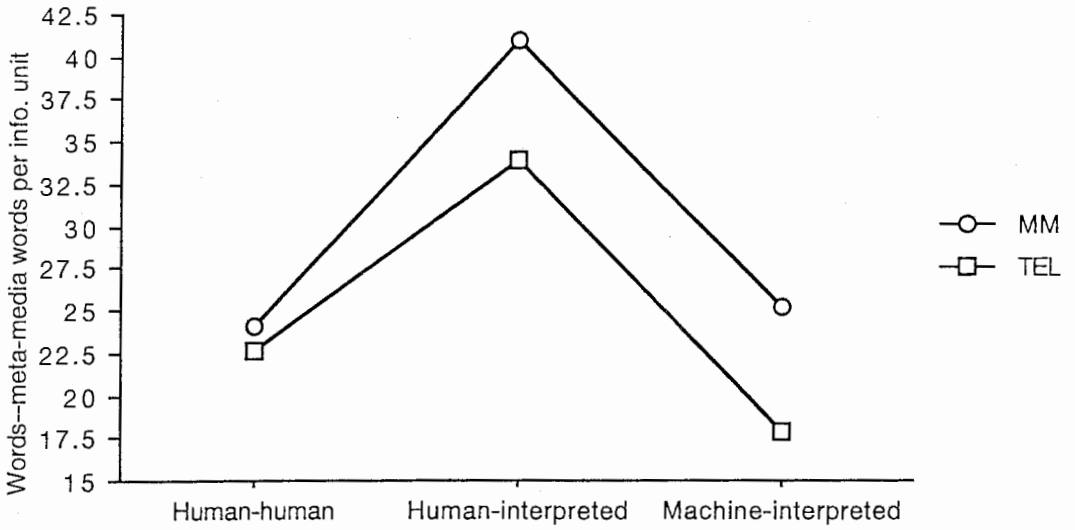

Figure 7. Words per information unit (with meta-media words subtracted out) for telephone and multimedia settings for all three experiments (differences are not significant).

Discussion

Do we need to re-think the old adage? It would appear so. Comparing experimental results across the telephone and multimedia settings reveals that subjects use about the same number of words to convey information whether they are communicating by telephone or via multimedia. (And this result is achieved only by ignoring the meta-media conversation that typically accompanies dialogue in the multimedia setting.) Subjects do not, in fact, use visual images to replace speech.

If we think about the everyday use of visual images, we realize that this is a reasonable result. It is rare that we allow images to *replace* words in everyday life. Newspaper, magazine, and book illustrations are invariably accompanied by captions; grandparents displaying pictures of their grandchildren never allow the picture alone to carry the message.

These anecdotal observations are supported by experimental evidence. In analyzing the drawings made by agent and client in the direction-finding task in these experiments, we found that these visual gestures were of two types. The first type, what we call "deictic" gestures accompanied a deictic expression in the speech of the subject, and did, indeed, convey information about that expression in a visual rather than verbal fashion. These gestures did tend to replace longer verbal descriptions. They accounted for only half the visual gestures used, however. The other half were what we call "redundant" gestures. These accompanied speech that contained no deictic expression and were simply visual correlates of the information that was *also* being expressed verbally. While the use of deictic gestures seems to support the adage, the use of redundant gestures contradicts it. In the latter case, subjects express information verbally despite the fact that the information is available visually.

The experimental design allowed us to examine the balance of speech and visual gestures across interpreting conditions as well. Doing so pointed up some revealing differences. In the human-human interaction, the presence of additional meta-media conversation alone accounted for the greater number of words in the multimedia setting. There was only slightly more information exchanged in that setting and analyzing words-per-information-unit had little effect on the relationship between the results for the telephone and multimedia settings.

However, in the interpreted conditions, the results were different. Note that the use of gestures increased in these settings (Figure 2). At the same time, meta-media conversation itself was not enough to account for the greater number of words; instead, it was necessary to consider the additional amount of information conveyed in the multimedia setting of these conditions. This suggests that in cases where the communication process is complicated by interpretation, subjects make greater use of visual gestures, and convey slightly greater amounts of information. It is only by considering the interaction of meta-media conversation and higher levels of information in the multimedia setting that the words used in that setting and in the telephone setting become equivalent.

Thus, two major results are evident. First, subjects do not use a smaller number of words in the MM setting. We cannot expect to reduce the burden on a language processing system simply by incorporating a MM component. Second, the MM setting is not a more efficient communication environment, if measured in terms of words per information unit. At best, that is, by ignoring meta-media conversation, it is as efficient as the telephone setting.

Users' Attitudes Toward Multimedia Communication[3]:

Motivation

The results discussed above are illuminated by the findings from post-experiment interviews. Clients were asked to rate their impressions along provided scales by marking an "X." The sum of their responses is represented collectively below.

Results

1. How would you rate how much you enjoyed the experiment?

Telephone:

a real bore	*kind of interesting*	*fun*	*had a great time*
	XXXXXX	X	X

Multi-media setting:

a real bore	*kind of interesting*	*fun*	*had a great time*
X	XXX	X	XXXXX

2. How would you rate how easy it was?

Telephone:

simple	*some effort*	*had to work at it*	*difficult*
XXXX	XXXXXX		

Multi-media set-up:

simple	*some effort*	*had to work at it*	*difficult*
XXXXXX	XXXX		

3. How would you rate the usefulness of:

Telephone:

very useful	*served some purpose*	*an inconvenience*	*worthless*
XXX	XXXXXXX		

[3]This description appeared in Fais (1994b).

Map:
_XXXXXXXX_____XX_____
very useful served some purpose an inconvenience worthless

Keyboard:
__X_____XXXXX____ *very useful served some purpose an inconvenience worthless*

In these interviews, subjects uniformly reacted in a positive way to the multimedia setting (Fais, 1994b; Park *et al.*, 1994). They reported greater enjoyment in the MM setting than in the telephone setting; novelty clearly plays a role in this reaction. However, subjects also felt that the MM environment was no harder to use, and in fact, might have been slightly easier. This seems to indicate that they, in fact, felt more comfortable in the MM setting than in the telephone setting. They cited the presence of the map and the capability of seeing directions marked on the map as having a positive influence on their ability to understand those directions and on their enjoyment of the task.

Discussion

Human beings vary in their ability to absorb information through visual, auditory and tactile channels, and strong visual learners, especially, appreciate the presence of the visual medium in a direction-finding task such as this. Subjects' greater confidence in and enjoyment of the multimedia setting is probably correlated with the increased amount of information conveyed in that setting.

Thus, despite the fact that the presence of the visual channel seems to have had little effect on reducing the amount of speech and thus the processing burden on an automatic speech processing system, it is still a worthwhile addition to such a system by virtue of the benefits it offers to the understanding and enjoyment of users, and to their ability to convey more information. While these results do not show clear advantages for the use of MM options in a language processing system, they do show some general advantages of the use of MM over the telephone, namely, greater information conveyed and greater ease and enjoyment of use.

Disfluency[4]

Motivation

We now turn to a more specific examination of the possible interactions between the MM communication environment and an automatic language processing system. If we can show that disfluency, a major obstacle both to speech recognition and to language processing, is lower in the MM setting, we can argue that the integration of a MM interface with a language processing system will improve the performance of that system.

Definition

Disfluency is typically considered to include false starts and filled pauses. These were transcribed for all conversations in each of the three experimental conditions. Disfluency rates were determined by calculating the number of false starts and filled pauses per 100 words.

Results

Disfluency rates for the agents do not necessarily reflect significant aspects of the experimental conditions. Rates for the agent in the human-human condition are those for

[4]Some of this analysis appeared in Fais (1994b).

native speakers of English; rates for the interpreter in the human-interpreted case are those for fluent, non-native speakers of English. While the (English-speaking) Wizard was a native speaker of English, his speech was carefully controlled to yield a zero level for disfluencies. (Figure 8.) Thus, in the discussion below, only the disfluency rates of clients will be considered.

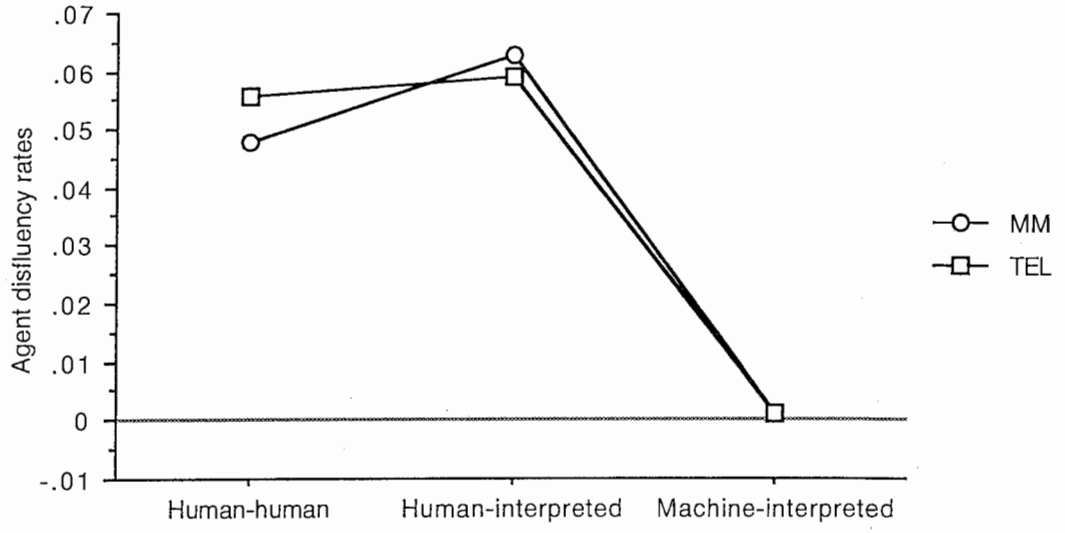

Figure 8. Disfluency rates for the agent, human-interpreter and Wizard in both MM and telephone settings.

Recall that each subject performed the tasks of the experiment in both telephone and MM settings. Subjects were evenly divided between telephone-first and MM-first orders. We can get an idea of the disfluency rates typical for each setting in each condition by looking at the results for that setting when it came first (Figure 9). Although the difference in disfluency rates in the two settings of the first experiment appears to be large, it is not, in fact, a significant difference. Thus, clients do not differ in disfluency rate depending upon setting.

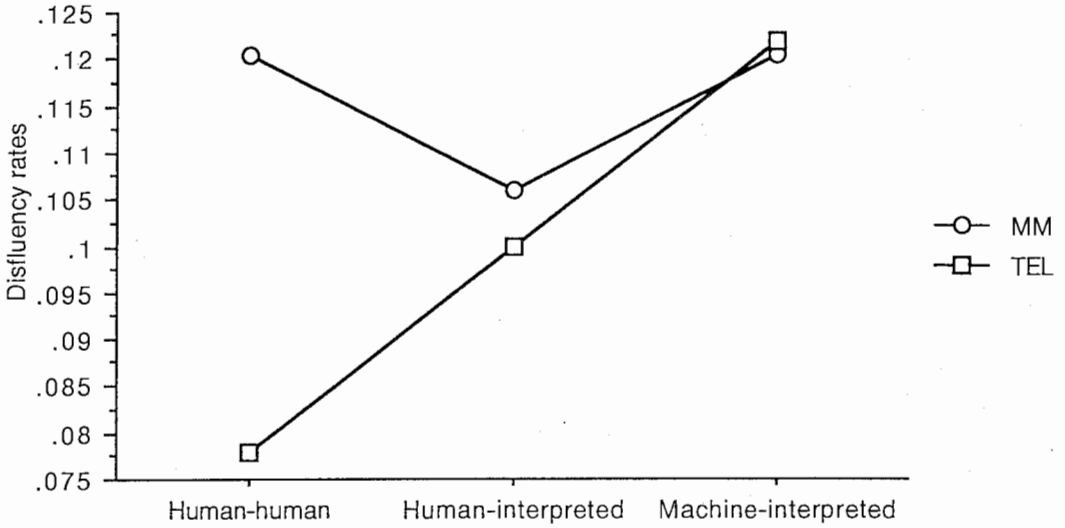

Figure 9. Client disfluency rates in setting of first trial only, for all three experiments.

However, there was an interesting effect of order of setting on disfluency rates. Disfluency rates for setting interacting with order were significant in the first, human-

human experiment; the subjects' speech tended to deteriorate in the second trial to a disfluency level slightly higher than that of the first trial. Settings did not show stable disfluency rates; rather, the disfluency rate in the second trial was based upon the rate of the setting of the first trial (Figure 10).

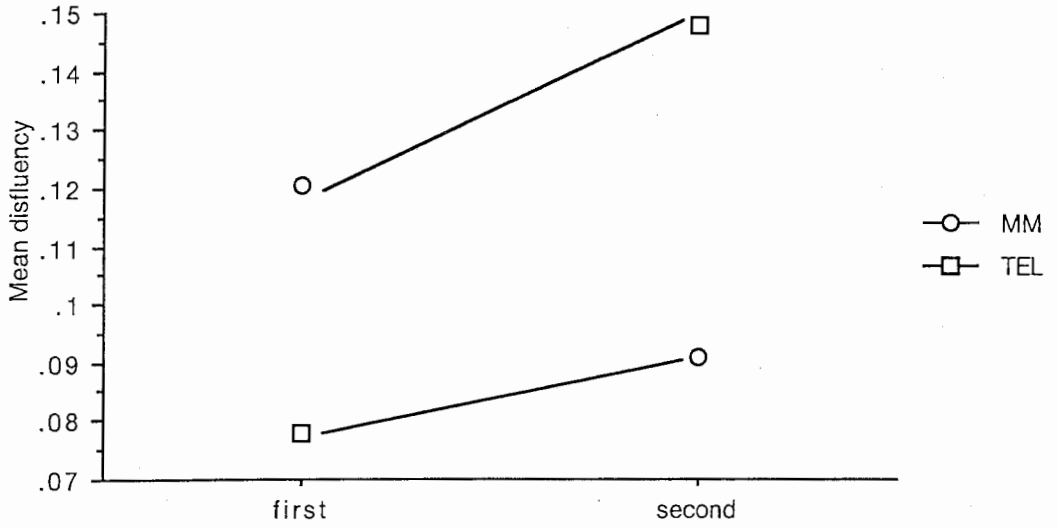

Figure 10. Client disfluency with respect to setting and order; human-human experiment.

This interaction was not significant in the human-interpreted or machine-interpreted cases (Figures 11, 12). However, it is interesting to note that there was a strong trend in the machine-interpreted case for disfluency rates to improve in the second trial, rather than to deteriorate, as they did in the human-human experiment (Figure 11).

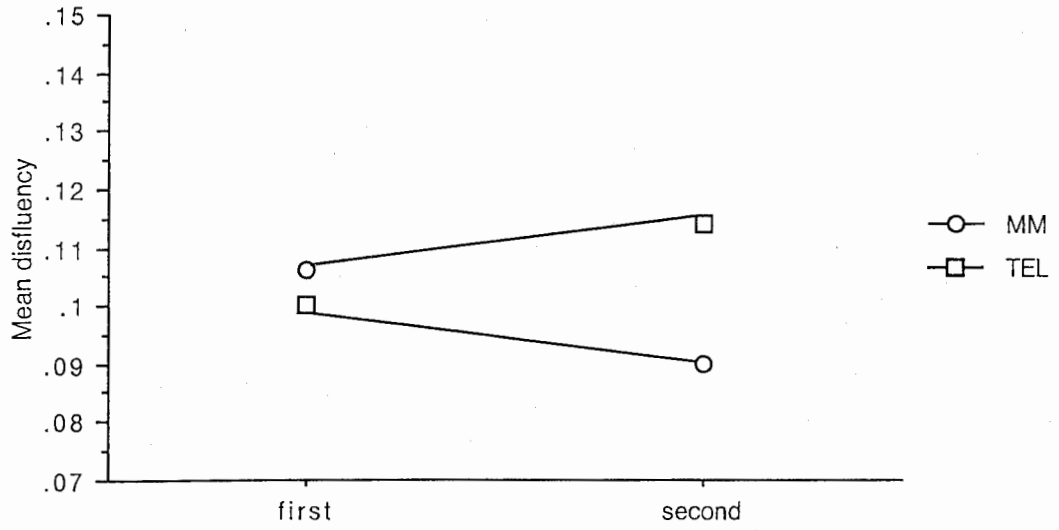

Figure 11. Client disfluency with respect to setting and order; human-interpreted experiment.

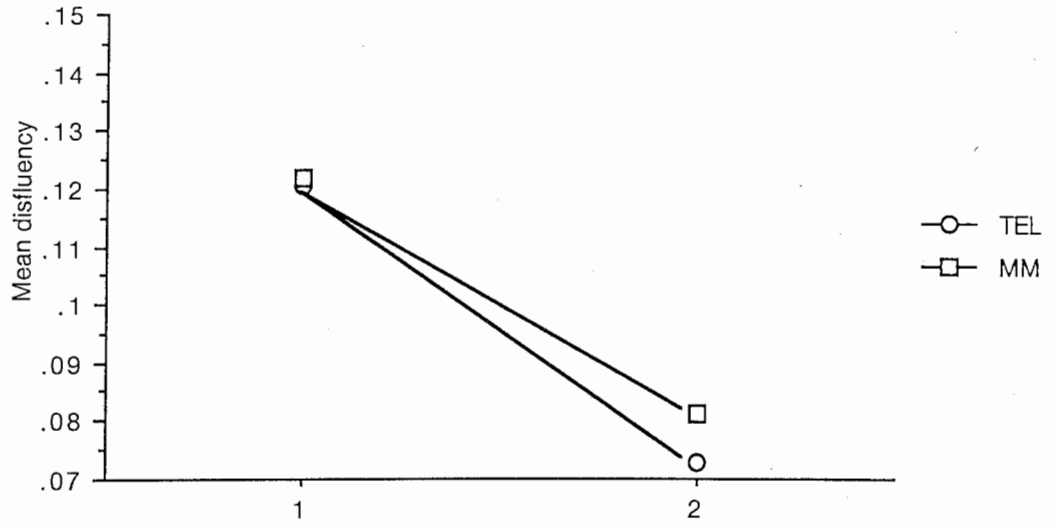

Figure 12. Client disfluency with respect to setting and order; machine-interpreted experiment.

Thus, while the setting of the interaction did not affect disfluency rate, it is encouraging to note that subjects, over time, became less disfluent in the machine-interpreted condition. This suggests that subjects try to "clean up their speech" in a human-machine interaction. In fact, disfluency rates in the *second* trials of each experiment were lowest for the machine-interpreted condition (Figure 13).

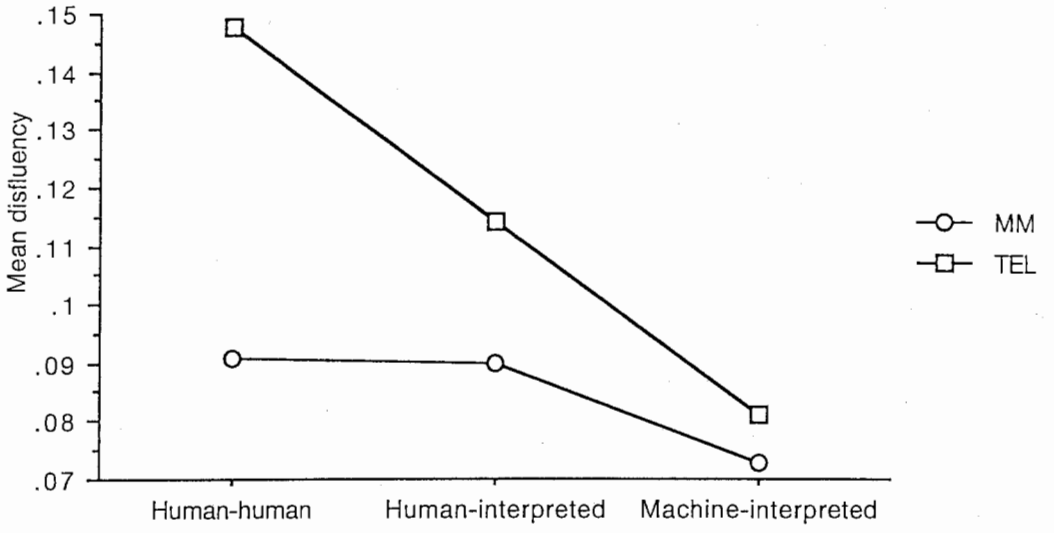

Figure 13. Client disfluency in the second setting for all three experiments.

Discussion

Disfluency rates do not differ between the telephone and MM settings. Thus, they do not constitute a good criterion for comparing the two. However, it seems clear from Figure 13, that clients, with practice, do exhibit a lower disfluency rate in the machine-mediated condition. This is consistent with findings of Suhm *et al.* (1994). In fact, with only two trials of practice, subjects' disfluency rates were lower in the machine-interpreted condition than in either of the other two conditions. Though this does not shed light on the differences between naive linguistic behavior in the MM and telephone settings, it is hopeful for the possibility of machine-mediated interaction.

Lexical Accommodation in Machine-Mediated Interactions[5]

Motivation

We focused on lexical accommodation in the MM setting as another possible way to enhance the performance of a language processing system. If users significantly adopt the lexical items used by their partners in cooperative dialogue, this information can be used to improve a language model incorporated in the processing system.

Introduction

For real-time, real-situation, human-computer interaction to approach reality, the burden of understanding and conveying information cannot be shared equally between the two interactors. Humans need to make allowances for features of the computer interface such as synthesized speech, limitations on the range of knowledge base, and imperfect speech recognition. However, in order for laymen to accept and use computers effectively in an interactive format, the restrictions placed upon users need to be as minimal and as natural as possible. For this reason, it is important to explore the linguistic behavior of humans interacting with computers in an unrestricted environment. In this way, it is possible to determine how humans are naturally inclined to accommodate to the current limitations of human-machine interaction. Encouraging those natural inclinations, then, in real human-machine systems will have a better chance of success than imposing artificial restrictions. In addition, systems designers can take advantage of the accommodations that humans make naturally to improve the performance of their systems (Fais *et al.*, 1995)

Below, we will discuss a particular kind of accommodation in conversational interaction, namely lexical accommodation. In lexical accommodation, one conversant adopts the lexical items used by the other conversant. This type of accommodation is one way in which users adapt to the limitations of computer interfaces, i.e., they converge to the limited lexicon of the computer. Thus, it has important implications for the design of workable human-computer interfaces. We will discuss results from the experiments briefly describe above: human-human monolingual, human-interpreted bilingual, and machine-interpreted bilingual.

Background

Of course, there is no *a priori* reason why interactors could not conduct conversations in completely different styles, using different phonologies, sentence structures, vocabulary, etc. However, it has been widely demonstrated that they do not. Lexical accommodation is only one instance of a wide range of convergence behaviors that humans display in conversation. Giles *et al.* (1987) cite studies demonstrating convergence of speech styles, dialect, non-verbal behavior, vocal intensity, prosody, speech rate and duration, and pause length. Garrod and Doherty (1994) report on a study in which conversational interactors in a language community showed a high level of convergence on a particular description language over the course of the task (a maze game). Fais (1994a) discusses both lexical and syntactic accommodation over a range of natural conversational styles.

Accommodation has also been shown to be present in human-computer interactions. Zoltan-Ford (1991) and Leiser (1989) demonstrate accommodation by users to the phrasing and vocabulary of the confirmation output of a computer in information manipulation and retrieval tasks. Speakers also unconsciously adapt their speaking behavior to the limitations of a speech recognition system, as demonstrated in Mellor and O'Connor (1995).

But while the phenomenon of accommodation, both between humans and between human and computer, is amply demonstrated, the motivations behind the phenomenon are less

---

[5]This analysis first appeared in Fais and Loken-Kim (1995b).

often discussed. Those in the field of human-computer interactions usually note simply that accommodation exists. They express relief when it is found to act as a natural constraint on the user's vocabulary or syntax (e.g., Leiser, 1989; Zoltan-Ford, 1991), and distress when accommodation to the "natural speech" style of some computer output encourages matching casual speech from the user which is difficult to process (e.g., Spitz *et al.*, 1991).

Speech Accommodation Theorists, who fall under a broad category which might be called socio-linguistics, tend to ascribe one of three motivations to speakers who accommodate: "evoking listeners' social approval, attaining communicational efficiency between interactors, and maintaining positive social identities." (Giles *et al.*, 1987) p.15). These motivations can be grouped into two major categories: concern with social standing and identity, and concern with communicational efficiency. Comparing human-human interactions with human-interpreted interactions allows us to gauge the importance of concern for communicational efficiency while that for social standing remains constant. Comparing human-interpreted and machine-interpreted interactions reveals the impact of concern for social standing while concern for communication efficiency remains the same.

Hypotheses. More specifically, if we base our predictions on the standard accounts in the literature, we should expect the following results concerning both level and direction of accommodation:

- In human-human interaction, we should find significant lexical accommodation. Because this is essentially an information-giving and -receiving task, we expect that the receiver of the information will accommodate to the giver, adopting the lexical items used by the speaker who imparts information.

- The human-interpreted setting constitutes both a human-human interaction and a more stressful communication environment, one in which communicational efficiency is a concern. For that reason, we expect an even greater level of accommodation in the human-interpreted setting than in the human-human setting. In the human-interpreted setting, we examine the accommodation between client and interpreter. The client is information receiver, and the interpreter is the imparter of information, not the originator; thus, neither client nor interpreter is in a dominant role. For this reason, we cannot predict whether the client will accommodate to the interpreter or *vice versa*.

- The machine-interpreted setting only indirectly involves human-human interaction; all dialogue is mediated by the "machine" interpreter. Therefore, we conjecture that interactors in this setting will not be concerned with social standing. On the other hand, this is the most difficult communication environment of the three, involving as it does not only the limited understanding of the machine translator but also limited speech recognition, a difficult-to-understand modulated speech signal, and rigid turn-taking constraints. For this reason, communicational efficiency *will* be a concern. However, whether this will generate more or less accommodation than concern for social standing generates in the human-human case is an open question.

- Since users in the machine-interpreted setting should not be concerned with social standing, we might predict a lower rate of accommodation than the human-interpreted setting. However, again, the greater difficulties in communication in the machine-interpreted setting might make up for a lack of concern with social standing, resulting in a rate of accommodation comparable to that in the human-interpreted setting. Results discussed below will shed some light on the interaction of these two factors.

- We expect that clients will accommodate to the machine to some extent, but that clients' word choice will also be affected by their perception of "what works," or "what the machine knows." In the Wizard-of-Oz situation used here, the vocabulary of the Wizards was not regulated. Their attempts to mimic computer behavior may have limited their word choice. Given this possibility, then, we predict that the

results will show client accommodation to the machine-interpreter, but at a lower level than in the human-interpreted setting.

Measures

Lexical accommodation. We measured lexical accommodation by examining the number of lexical items which were used by *both* interactors in the course of a conversation. The accommodation rate for each conversation in the three experiments was calculated by dividing the number of (different) lexical items the two speakers had in common by the total number of (different) lexical items in the conversation.

We calculated lexical accommodation rates for client and agent in the same-language, human-human experiment setting; for client and (Japanese-to-English) interpreter in the bilingual, human-interpreted experiment setting; and for client and (Japanese-to-English) "Wizard" interpreter in the "machine-interpreted" experiment setting. Although the actual measurement of lexical items was done for the English speech of the Japanese-to-English *interpreters*, these interactors in the conversation will be referred to below as "agents," to conform to the human-human setting in which we assessed the lexical accommodation of the agents directly.

Direction of accommodation. Accommodation is not necessarily a mutual phenomenon (Giles, 1987). In order to determine if one conversant was accommodating more than the other, we examined the number of words used first by the client and the number used first by the agent. We reasoned that the subject who used a particular lexical item first was *not* accommodating, and, by extension, then, the subject who did not use an item first, *was* accommodating.

The following objection to this definition might be made: the fact that interactors use words that their partners have used does not necessarily mean that they are accommodating to the other's prior use of that word. But what other justification could there be for saying that accommodation *is* taking place? Because even lexical accommodation is rarely a conscious act, speakers' intuitive judgments are not helpful. On the other hand, outside observers have no external evidence on which to base such a judgment. Thus, we will use the quantitative criterion described above. We will argue that accommodation is a real phenomenon in dialogue; it follows, then, that at least some of the instances in which conversants use lexical items previously used by their partners are instances of accommodation.

Relative importance of words-in-common. We also wanted to look beyond the initial use of lexical items to determine what role was played in subsequent conversation by the words that agent and client both used. That is, once one conversant accommodates to the other by adopting a lexical item, does that conversant continue to use that lexical item in a significant way in the remainder of the conversation? In order to explore this question, we estimated the percent of usage for each word-in-common, for both client and agent. That is, for each word-in-common, we divided the number of times each subject used the word by the total number of words uttered by that subject in order to determine what proportion of the subject's conversation consisted of the uses of that word. By comparing these proportions for the roles of client and agent, we ascertained the relative frequency of the word for each role, giving us some idea of the "importance" of the word for that role.

Coincidental overlap. Of course, a certain amount of lexical overlap is inevitable as a simple artifact of cooperative conversation. In order to determine the extent of coincidental overlap, we measured the lexical overlap in the speech of clients and agents from the first experiment who had not participated in the experiment together. That is, the speech of clients who had participated in the experiment with Agent A was compared with the speech of Agent B. Likewise, the speech of clients who had participated in the experiment with Agent B was compared with the speech of Agent A. Because these conversants were not talking to one another, the lexical overlap in their speech could not

be a result of accommodation to one another. However, because the overlap was calculated for speakers engaging in cooperative dialogues concerned with the same task and via the same media (telephone and multimedia), it reflects the extent to which overlap occurs simply because these are speakers in similar situations talking about similar topics.

Results

None of the measures we examined showed significant differences with respect to setting. That is, MM was neither more nor less effective vis-à-vis lexical accommodation than the telephone setting. However, the results reported below *did* reveal certain important characteristics of the interpreting conditions investigated.

Lexical accommodation. The rates for lexical accommodation for all three experiments were significantly different from the level established for coincidental lexical overlap (Figure 14, Table I; data were subjected to two-way analyses of variance). In addition, the lexical accommodation rates for each experiment also differed significantly from those for each of the other two experiments. Human-human accommodation was higher than coincidental overlap, but lower than both of the interpreted settings. The human-interpreted setting had the highest rate of accommodation.

Table I. Significance levels for differences in lexical accommodation (two-way ANOVA).

	human-human	human-interpreted	machine-interpreted
Coincidental overlap	p<.03	p<.0001	p<.001
Human-human		p<.0001	p<.02
Human-interpreted			p<.01

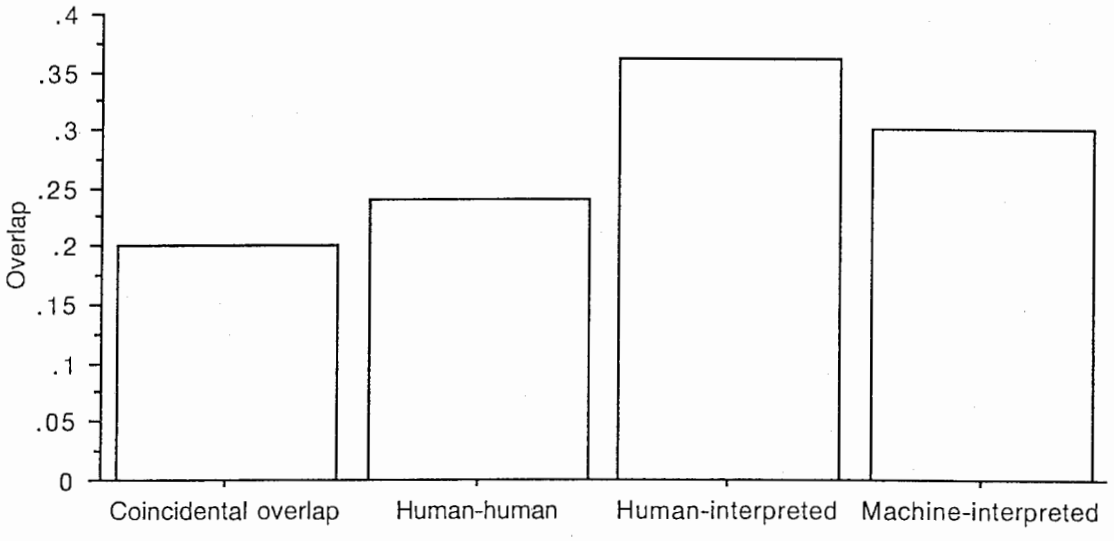

Figure 14. Rates of accommodation for coincidental overlap and all three conditions.

Direction of accommodation. When we examined the percentage of words-in-common used first by each role (agent or client), the following patterns emerged (Figure 15; data were subjected to three-way analyses of variance). In the human-human setting the agent used a significantly higher percentage of words first (p<.03); the client accommodated to the agent. In the human-interpreted setting, there was no difference. It is not possible to say that one or the other interlocutor was responsible for the accommodation found. In the machine-interpreted setting, the agent used a significantly higher percentage of words first (p<.005); the client accommodated to the agent.

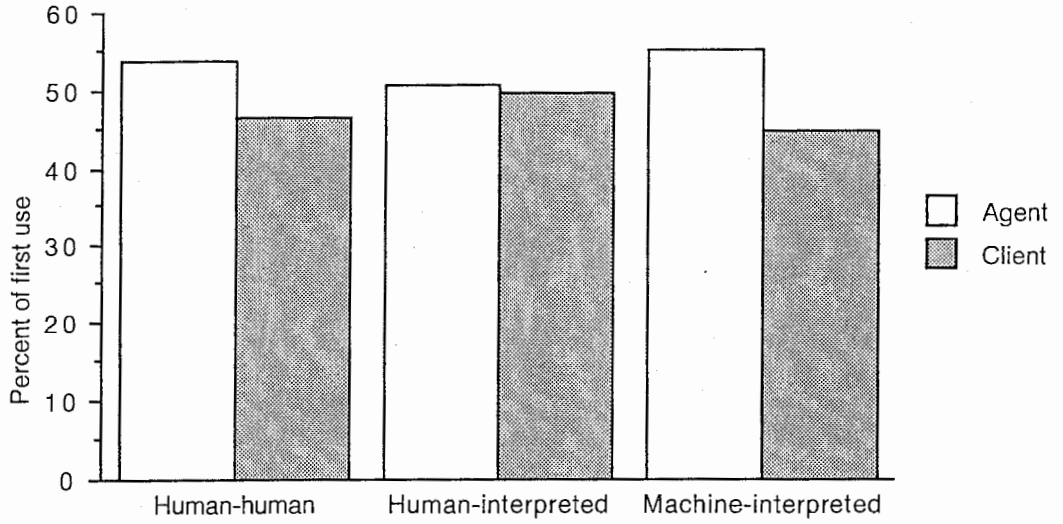

Figure 15. Percent of first use of words-in-common for agent and client in each setting.

Relative importance of words-in-common. An examination of the use of each word-in-common with respect to overall word use for client and agent, i.e., the word's "importance," showed the following results (Figure 16; three-way ANOVA). In the human-human setting, client use of words-in-common made up a significantly greater percent of total word use than agent use of these words ($p<.0003$). There was no significant difference between client and agent in the human-interpreted setting or in the machine-interpreted setting.

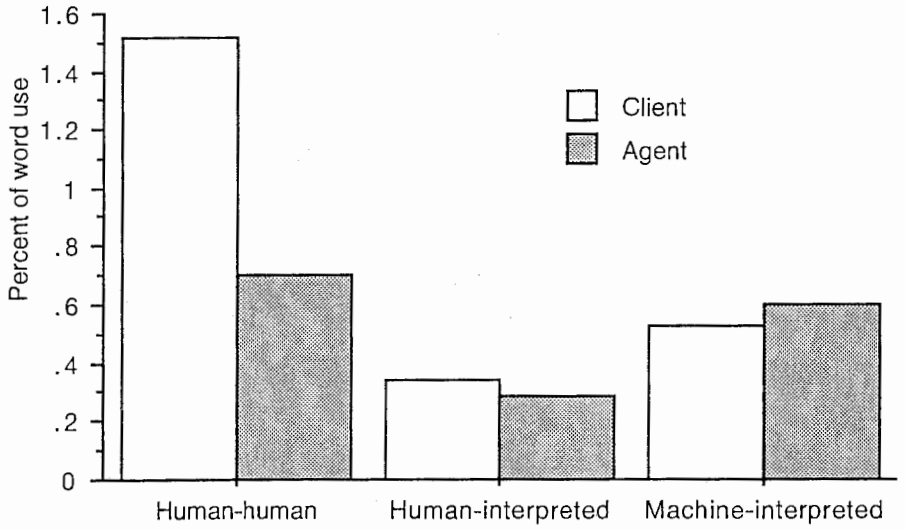

Figure 16. Frequency of use of words-in-common for agent and client in each setting.

Discussion

We analyzed lexical accommodation in a variety of interactions in order to determine how accommodation can be expected to operate in a machine-interpreted context, and learn ways in which to support lexical accommodation in the design of human-machine interfaces. It is encouraging that lexical accommodation happens spontaneously. As our initial results show, it is not simply a coincidental byproduct of conversing about common topics. There is a significant difference between that case (what we have called "coincidental overlap") and the case of two people talking about the same topic *to one another*. While lexical accommodation has been shown for typed human-computer interactions (Leiser, 1989), and lexical and structural accommodation has been

demonstrated for typed and spoken human-computer interactions in constrained contexts (Zoltan-Ford, 1991), these results demonstrate lexical accommodation for unconstrained spoken human-computer interactions. Given that lexical accommodation is a "real" phenomenon, then, how can we characterize the patterns of accommodation, and what can we learn from them?

It is important to note in the context of an evaluation of the MM interface, that there were no significant differences between results for the MM and telephone settings. This shows that the interface involved in these conversations does not affect the trend of the results discussed below.

In the human-human setting, there was a non-trivial, but low level of accommodation. The client accommodated to the agent, using the words-in-common more frequently in subsequent conversation than did the agent. This is consistent with the interpretation that the agent acted as information-provider and the client acted as information-receiver in a non-stressful communication environment, as our initial hypothesis stated. The interactors maintained a level of accommodation high enough to satisfy their concern for social standing, but since they were native speakers of the same language and the communication channel was clear and direct, the interaction was relatively stress-free and straightforward. They had minimal concern for communicational efficiency, and there was no incentive to extend lexical accommodation. This interpretation is confirmed by the client's tendency to use the accommodated lexical items to a higher degree than the agent in the course of the conversation. The use of these lexical items may have been one way in which the client signaled his understanding and reception of the information from the agent.

On the other hand, the human-interpreted setting presented a more difficult communication environment in which concern for communicational efficiency was present. Since the interaction was also human-human, social standing continued to be a concern. We expected that the addition of the concern for efficiency to that for social position would result in a higher level of accommodation; in fact, the level of accommodation observed in the human-interpreted setting was the highest in all three experimental settings.

More specifically, the human-interpreted setting involved speakers from two different language backgrounds, both of whom were capable of recognizing the differences in their linguistic behaviors, and of reducing those differences to facilitate communication. Lexical choice is a surface level phenomenon, open to manipulation. Thus, in addition to signaling understanding as in the same-language setting above, lexical accommodation is an important conversational strategy for speakers who do not share linguistic conventions. The interpreters in the human-interpreted setting were native speakers of Japanese, and, while fluent in English, the range of overlap between their English linguistic habits and those of the native English-speaking clients was much smaller than that between two native speakers. Lexical choice was an obvious strategy for establishing shared linguistic behavior, and thus promoting effective communication. So, concern for social standing and communicative efficiency combined to encourage a high rate of mutual accommodation.

Because neither client nor interpreter had a dominant role in the conversation, we could not predict the direction of accommodation. In fact, our results show that it was not possible to single out a primary accommodator in the human-interpreted setting, either in terms of proportion of words used first or the frequency with which words-in-common were used. Does this mean that both speakers accommodated or that neither did? Considering the high accommodation rates for this setting (Figure 14), we conclude that, in fact, *both* client and agent were accommodating to one another.

In the machine-interpreted setting, we saw a rate of accommodation higher than that of the human-human setting, but lower than that of the human-interpreted setting, as expected.

The machine-interpreted setting is probably even more stressful a communication environment than the human-interpreted setting; concern for communicational efficiency resulted in a higher level of accommodation than concern for social standing did in the human-human setting. However, we do not expect humans to be concerned about their social standing with a machine, unlike in the human-interpreted setting. This explains why the rate for lexical accommodation in the machine-interpreted setting is lower than that of the human-interpreted setting. The greater concern for communicational efficiency in the machine-interpreted setting was not enough to generate as high a level of accommodation as that found in the human-interpreted setting, where there was the additional factor of concern for social position, though it did generate a higher level of accommodation than did concern for social standing alone (the human-human setting).

As we conjectured above, clients accommodated to the machine as part of a strategy for effective communication. However, given the fact that there was a lower rate of accommodation than in the human-interpreted setting, coupled with the strong directionality observed, we conclude that this is not a case of mutual accommodation. Instead, as in the human-human setting, clients were the primary accommodators. Clients in the machine-interpreted setting may have perceived the machine to be in the dominant role, just as the agent played the dominant role in the human-human setting.

Recall that clients in the machine-interpreted setting, unlike those in the human-interpreted setting, did not use words-in-common more than the agent did in subsequent conversation. Since clients were not concerned with social standing, including signaling understanding and establishing mutual linguistic conventions, accommodation in the machine-interpreted setting was a local phenomenon which did not extend throughout the conversation.

Conclusion

What does this tell us about the design of human-computer interfaces? Recall that these conversations were unconstrained; neither agents, clients nor interpreters, whether human or machine, were under instructions to limit or modify their speech in any way. Thus, what we see in these results is the natural tendency of humans to accommodate to their interlocutors in a variety of communication environments. This tendency resulted in the highest level of accommodation in the human-interpreted setting. That level was achieved as a result of mutual accommodation between the two humans involved, both of whom felt a concern for both social standing and communicational efficiency. The level of accommodation observed in the machine-interpreted setting was both lower and less extensive, i.e., it did not persist throughout the conversation.

We can take advantage of even the moderately high level of accommodation found in the machine-mediated setting by building into a language processing system a preference for the lexical items used by the machine. This could improve the recognition and parsing of the natural speech of users at least within a small range. Coupled with accommodations in other aspects of language such as discourse and syntactic complexity, fluency, and speaking rate (Fais *et al.*, in press) lexical accommodation can inform a language model to improve language processing performance by exploiting the relationships between human speech and the speech of the machine interface.

We would also like to investigate the possibility of increasing the level of accommodation in the machine-mediated setting to the level found in the human-interpreted setting. Ideally, we would like users' accommodation to a machine interface to be as high as possible so that the lexical variability of users' speech can be as constrained as possible. In this analysis, the results obtained in the MM and in the telephone settings were equivalent; however, the MM interface has greater potential for increasing lexical accommodation. The resources of a multimedia environment can be used to replicate the effect of the human-interpreted setting by providing the machine interface with a human-like persona. A number of human-computer systems already include such a feature (e.g., Ball and Ling, 1995; Bertenstam *et al.*, 1995; Webber, 1995); it remains to be seen if it

will have the desired effect on lexical accommodation. On the other hand, there is evidence that encouraging users to interact with machines as if they were humans may actually undermine the quality of the users' speech from the point of view of language processing. Work in the area of disfluencies in human-to-machine speech suggests that humans do, in fact, "clean up" their speech for machines (Suhm *et al.,* 1994; and above). These advantages may be lost if humans are encouraged to treat a machine interface as if it were human. Empirical investigation is required to determine if an optimal balance can be reached.

We have suggested that the design of speech recognition and language processing systems can take advantage of users' lexical accommodation to machine interfaces to improve system performance. This, in turn, would allow the construction of systems which make fewer demands on the willingness of users to adjust to misrecognitions and misunderstandings. Enhancing computer interfaces with multimedia-generated "persona" could encourage users to interact with computer interpreters as if they were interacting with human interpreters. This result would also have the effect of further increasing lexical accommodation from users.

Summary

When we attempt to compare the effectiveness of the telephone and MM settings, the results discussed above are equivocal. With respect to information exchange, users tend to convey more information in the MM setting. On the other hand, they use more words to do so. The presence of meta-media conversation seems to be a disadvantage in the MM setting. However, when we look at alternative criteria such as ease of use, confidence in information received, and enjoyment, the MM setting is rated much more highly than the telephone setting.

Looking at the advantages of integrating a MM system with an automatic language processing system, we still find no differences between the telephone and MM settings. However, we do see some positive results. Subjects do utter fewer disfluencies in the machine-interpreted condition than in any other condition, with practice (that is, in their second trials). We find an elevated level of accommodation in this condition, as well, though this level is not quite as high as that in the human-interpreted condition. There are indications that using a MM interface to full advantage could raise the lexical accommodation level.

These rather equivocal results suggest that the configuration of MM options and language processing in these experiments was not optimal. These results suggest some directions for improvement, however. One major focus is meta-media conversation. Clearly an effective integration of MM and language processing would be one in which subjects felt comfortable enough with the system that they did not need to discuss the system itself to the extent that they did in these experiments. The analysis of accommodation suggests that "personifying" the processing system might raise accommodation levels (though care needs to be taken that it does not at the same time raise disfluency levels).

Protocol "Experiment"

Motivation

From post-experiment interviews and the analysis of the results of the initial experiments as discussed above, it became clear that there were two factors which might skew the behavior of subjects using the EMMI environment. The first was the fact that this environment was novel; subjects did not always clearly understand what its potential was or how to make the most of it. The second, related, factor was that subjects were often intimidated by the technology: not only were they uncertain about how to use it, they also were hesitant to do so. They often checked the appropriateness of their behavior with their conversational partner, with such meta-media utterances as "I'm going to type now" or "Can you see this on the map?". High levels of meta-media conversation in the MM

setting confirm this interpretation. In light of these factors, subsequent research was planned to determine what configuration of technology and instruction would allow the most efficient use of the system.

Method

A protocol analysis was planned in order to determine how *expert* computer users, knowledgeable about the component technologies of EMMI, would proceed through the experimental tasks. The purpose of the procedure was to learn from these expert users the best way to instruct, train, and conduct goal-oriented conversations with subjects via the MM environment of EMMI. Specific areas addressed included: instructions before the experiment; instructions given by the wizard, i.e., the machine (e.g., "please rephrase," in case of misunderstanding); instructions given by the agent (e.g., "please type in your name," etc.); the need for some kind of feedback from the system to make the use of MM more efficient; the nature of the accomplishment of the task, whether it should be more machine-initiated, more form-based, etc.; and any other ideas from competent users about communicating via a MM setting that could help users to be as efficient as possible.

The "experiment" was conducted as an "open" WOZ experiment. That is, exactly the same set-up was used as for the previous WOZ experiment, but the subjects knew that the "Wizard" was translating their speech. Five researchers working in various areas of communication science at ATR participated as "clients." A Japanese-speaking "agent" and a Japanese/English translator acting as "Wizard" also participated in order to match as closely as possible the setting of the previous experiments. All "clients" had the system and the task explained to them and were asked to work through the task, commenting on the EMMI environment as they felt was appropriate. Their speech was recorded on DAT tape and their actions and discussion with the researcher conducting the analysis were videotaped. The tapes were later reviewed and all reactions noted.

Results

Participants made a large number of specific suggestions concerning the appearance and function of the multimodal interface. The major issues which seemed to recur throughout the users' responses are discussed here.

Translation. The "Wizard" in previous experiments had been instructed to wait for several seconds before commencing translation, in order to simulate an actual system working in not-quite-real time. All users expressed dissatisfaction with the slowness of the translation. Further, the voice distortion used to simulate computer speech was also judged to be poor. That is, although it was completely convincing to the layman, it did not accurately represent the current capabilities of synthesized speech. Therefore, the suggestion was made to pre-synthesize typical utterances used by the agent (such as "This is the International Conference Office; how can I help you?" and the like). These pre-prepared utterances could be linked to buttons which the Wizard could push at appropriate times in the conversation. This would speed up the "translation" and render it in actual synthesized speech. Some sentences could also be prepared in template form, and the Wizard could simply type in the variable information. An example might be: "The price for a [single/double] room in the [name] Hotel for [one, two...] night[s] is [amount]." For sentences not pre-prepared, the Wizard could simply type in the sentence and have it synthesized in real time.

Interaction with the system. In the current version of EMMI, there is no interaction possible between the client and the system itself. Users expressed the need to know what the system was doing at each stage in the translation process and to have some control over the system itself.

To address the first need, users suggested that the video window could be better utilized. That is, the video window could show the current talker, including an icon for the

translation system. A caption below the video could notify the user: "translating for the client," "listening to the agent" and so forth.

To address the second need, users suggested the creation of buttons to ask the system to repeat its last utterance, or to rephrase its last utterance, and to interrupt the system, stopping a translation.

Role of the Agent. It became clear in the course of the analysis, that a number of cases of confusion could be avoided if the agent took a more assertive role in managing the behavior of the client. Recommendations were made for the agent to: ask the client explicitly to type information on the hotel reservation form; type in Romaji all unfamiliar Japanese words spoken to the client; and take numbers and letters (spellings) with no intermediary translation. In addition, the agent should not have to ask the client his or her location; that information should be automatically available through the telecommunications system.

New System Design

Introduction

The Appendix contains schematic drawings of the new design of the MM computer screens for the client. Below, we discuss how the changes made to the system in each area identified by the expert users address the problems revealed in the experimental analyses discussed above.

Translation. The new EMMI system incorporates the changes outlined above as well as a number of more detailed changes in the configuration of the system. It also has a user interface so that the system can stand alone. Because the previous system was an experimental one, there was no provision for automatically training users; experimenters were always on hand to instruct and guide the users' practice. Currently, however, instruction and practice are incorporated into the system interface itself.

While the use of actual synthetic speech, (i.e. CHATR), does not directly address any of the concerns above relating to the efficiency of the system, it does allow us to move closer to our goal of a fully integrated multimedia automatic language processing system. As such, subsequent experiments in the new EMMI will allow us to collect data on the usability of CHATR.

Interaction with the system. One source of "extra" words in the machine-mediated condition may be the negotiation of turns. That is, with the lengthy time lapse between an utterance and its translation, subjects sometimes spoke too soon, contributing utterances out-of-turn that then confused the interaction and required further discussion or negotiation to resolve. The addition of a window for tracking the conversation addresses this problem (see the Appendix). This window contains the face of the agent, the "face" of the computer and the face of the client. When the computer is translating for or listening to the client, he faces the client with his mouth open or closed, respectively. When he translates for or listens to the agent, he faces the agent with mouth open or closed. In addition, a verbal description of the current state of the conversation is displayed. The intent of this addition to the system is to eliminate verbal queries concerning the state of the conversation, as well as misplaced utterances which may confuse the conversation, leading to additional, off-task words.

Clients can also control the system to some extent. They can push an "interrupt" button to stop a current translation, a "repeat" button to hear the last translation produced again, or a "rephrase" button to have the last translation reworded. These buttons are intended to replace verbal requests for repetition or expressions of misunderstanding.

Role of the agent. One problem with the human-interpreted and machine-interpreted experimental conditions was that they were conducted using naive agents. As a result,

these agents did not take the amount of initiative that trained agents in a realistic Conference Office scenario could take. Further work in the new system will employ a trained agent who can circumvent some meta-media questions by, for example, requesting a client from the start to type in certain information, or by offering explicitly to type for the client. The addition of a "You are Here" symbol on the map in the direction-finding task, also a more realistic feature, should eliminate negotiation about the client's initial position.

Additional changes. The new system is also more self-contained. That is, the client requires minimal training by the experimenter before he/she is able to use the system. Part of the reason for that is that the system is somewhat simplified; clients no longer have the options to change their pen color, clear their drawings or change the video window size, for instance. In addition, the system itself takes more initiative in getting the client started. The first several screens require minimal input on the part of the client (see the Appendix). Finally, instructions are included whenever an option is presented for the first time. These have the dual function of training the client in the use of that option and allowing the client to grow accustomed to the quality of the synthesized speech by listening to it and reading the output at the same time.

Future Work

The next step in our research must be to determine if, in fact, the changes made in the new system will be effective in increasing the efficiency of a MM setting integrated with automatic language processing. We will compare the behavior of users in the new system with that of users in the previous system. We hope that the changes made will reduce the amount of meta-media conversation, and thus the number of words used in the MM setting, so that the burden on the language processing system can be reduced. We hope to maintain the benefits seen in lower disfluency rates and in greater amounts of information exchanged. Further, we will test the hypothesis that the addition of a "persona" to the machine-mediated condition will raise the level of accommodation; at the same time, we will investigate whether the addition of that persona also degrades the fluency of the speech of the clients.

In our future work, we will also return to a point made in the analysis of our very first EMMI experiment; namely, that experience may play a role in subjects' behavior (Fais, 1994b). In that analysis, we surmised that clients may use more words and more meta-conversation in the MM setting because they are unfamiliar with that setting. In our first experiment in the new system, we will compare the behavior of expert (computer) users and relatively naive users of the new system to determine to what extent familiarity with computer technology affects subjects' behavior in the MM setting.

However, the somewhat inconclusive results from previous analyses suggest that the criteria we are using in order to measure the effectiveness of the system may not be the most revealing choices. For this reason, in future experimental work, in addition to the measures above, we will investigate in a more rigorous way the nature and amount of the information conveyed in each setting as well as the confidence subjects have in that information. We will also examine the relationship between the nature of the gestures used by the subjects and the linguistic information that accompanies these gestures in order to maximally exploit this interaction to design a workable system.

References

André, E. and T. Rist. 1993. The design of illustrated documents as a planning task. In *Intelligent Multimedia Interfaces*, M. Maybury, ed., pp. 94-116. Menlo Park, CA: AAAI Press/MIT Press.

Arens, Y., E. Hovy, and M. Vossers. 1993. On the knowledge underlying multimedia presentations. In *Intelligent Multimedia Interfaces*, M. Maybury, ed., pp. 280-306. Menlo Park, CA: AAAI Press/MIT Press.

Ball, J. Eugene, and Daniel T. Ling. 1995. Spoken language processing in the Persona Conversational Assistant. In *Proceedings of the ESCA Workshop on Spoken Dialogue Systems,* pages 109-112, Vigsø, Denmark, May-June.

Bertenstam, Johan, *et al.* 1995. The Waxholm system: A progress report. In *Proceedings of the ESCA Workshop on Spoken Dialogue Systems*, pages 81-84, Vigsø, Denmark, May-June.

Fais, Laurel. 1994a. Conversation as collaboration: Some syntactic evidence. *Speech Communication,* 15: 231-242.

Fais, Laurel. 1994b. Effects of communicative mode on spontaneous English speech. Technical Report of the Institute of Electronics, Information and Communication Engineers, NLC94-22(1994-10):1-8.

Fais, Laurel, and Kyung-ho Loken-Kim. 1995a. How many words is a picture really worth? Technical Report of the IEICE, SP95-81(1995-12) , The Institute of Electronics, Information and Communication Engineers.

Fais, Laurel, and Kyung-ho Loken-Kim. 1995b. Lexical Accommodation in Human-interpreted and Machine-interpreted Dual Language Interactions. In *Proceedings of the ESCA Workshop on Spoken Dialogue Systems,* Vigso, Denmark, May-June.

Fais, Laurel, Kyung-ho Loken-Kim, and Tsuyoshi Morimoto. (in press) Linguistic and para-linguistic differences between multimodal and telephone-only dialogues in English and Japanese. *J. Acoustical Society of Japan (E).*

Fais, Laurel, Kyung-ho Loken-Kim, and Young-Duk Park. 1995. Speakers' responses to requests for repetition in a multimedia cooperative dialogue. In *Proceedings of the International Conference on Cooperative Multimodal Communication*, pages 129-144, Eindhoven, The Netherlands, May.

Garrod, Simon, and Gwyneth Doherty. 1994. Conversation, coordination, and convention: An empirical investigation of how groups establish linguistic conventions. *Cognition*, 53:181-215.

Giles Howard, Anthony Mulac, James J. Bradac, and Patricia Johnson. 1987. Speech accommodation theory: The first decade and beyond, in M.L. McLaughlin, editor, *Communication Yearbook 10*, pages 13-48. Sage Publications, London, UK.

Leiser, R.G. 1989. Exploiting convergence to improve natural language understanding. *Interacting with Computers: The Interdisciplinary Journal of Human-Computer Interaction,* 1(3):284-298.

Loken-Kim, Kyung-ho, Fumihiro Yato, Kazuhiko Kurihara, Laurel Fais, and Ryo Furukawa. 1993. ATR Technical Report TR-IT-0018. ATR Interpreting Telecommunications Research Laboratories, Kyoto, Japan.

McCaffery, F., M. McTear, and M. Murphy. 1995. Designing a multimedia interface for operators assembling circuit boards. In *Proceedings of the International Conference on Cooperative Multimodal Communication,* (Eindhoven, The Netherlands, May 24-26, 1995), pp. 225-236.

Mellor, Brian, and Cian O'Connor. 1995. User adaptation to voice input interfaces. In *Proceedings of the ESCA Workshop on Spoken Dialogue Systems*, pages 97-100, Vigsø, Denmark, May-June.

Park, Y., K.H. Loken-Kim and L. Fais. 1994. An experiment for telephone versus multimedia multimodal interpretation: methods and subjects' behavior. ATR Technical Report TR-IT-0087, ATR Interpreting Telecommunications Research Laboratories, Kyoto, Japan.

Park, Y., K.H. Loken-Kim, L. Fais, and S. Mizunashi. 1995. Analysis of gesture behavior in a multimedia/multimodal interpreting experiment; human vs. Wizard of Oz interpretation method. ATR Technical Report TR-IT-0091, ATR Interpreting Telecommunications Research Laboratories, Kyoto, Japan.

Spitz, Judith. 1991. Collection and analysis of data from real users: Implications for speech recognition/understanding systems. In *Proceedings of the 4th Darpa Workshop on Speech and Natural Language,* pages 164-169, Asilomar, California, USA, February.

Suhm, B., L. Levin, N. Coccaro, J. Carbonell, K. Horiguchi, R. Isotani, A. Lavie, L. Mayfield, C. P. Rosé, C. Van Ess-Dykema, and A. Waibel. 1994. Speech-language integration in a multi-lingual speech translation system. In *Proceedings of the AAAI Workshop on Speech and Language Processing*, Seattle, Washington, USA.

Walker, M. 1992. Redundancy in collaborative dialogue. In *Proceedings of 14th Coling*, pp.345-351.

Webber, Bonnie. 1995. Instructing animated agents: Viewing language in behavioral terms. In *Proceedings of the International Conference on Cooperative Multimodal Communication*, pages 5-15, Eindhoven, The Netherlands, May.

Zoltan-Ford, Elizabeth. 1991. How to get people to say and type what computers can understand. *International Journal of Man-Machine Studies,* 34:527-547.

Appendix

1　はじめに

このドキュメントは、マルチモーダル実験の為のシュミレーターシステムの構成や動作などに関するドキュメントです。このシステムは、従来のシュミレーターに比べてより使いやすくするために作成されました。また、従来通訳者からクライアントへの音声は生の声を流していましたが、今回は音声合成を使用しました。

2　システム概要

2.1　システム構成

マシンは全部で4台使用し、その内訳は通訳者用に2台、エージェント用に1台、クライアント用に1台。
それぞれ、ソケットを通して通信します。
その通信を制御するプロセスをサーバーとして、各プロセスは必ずここを通してメッセージが行き来します。
詳しいことは、図1：システム構成を参照して下さい。

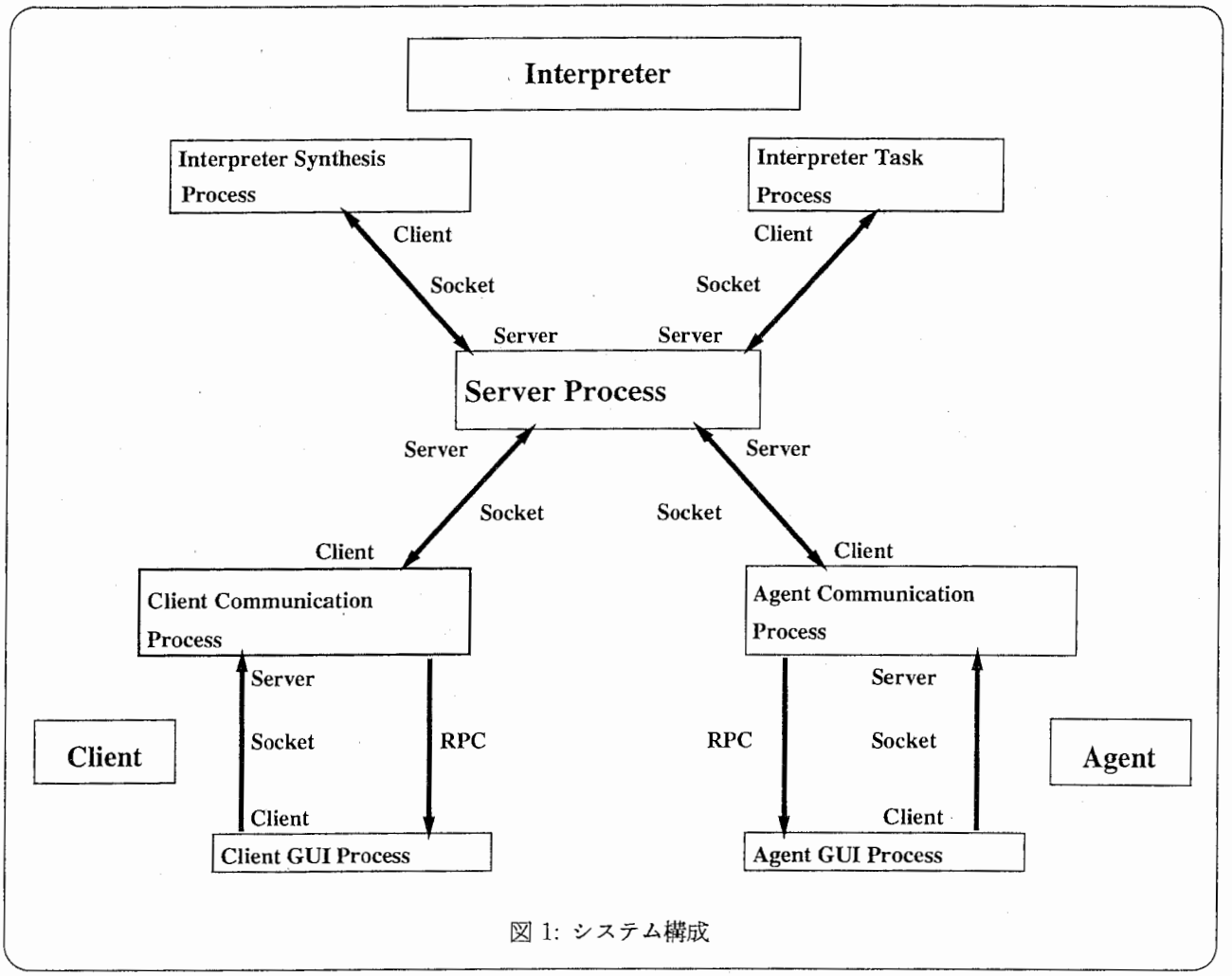

図 1: システム構成

2.2 エージェント用システム

エージェント用システムは、ＧＵＩ（Ｇｒａｐｈｉｃ　Ｕｓｅｒ　Ｉｎｔｅｒｆａｃｅ）プロセスと、サーバーとの通信をする通信用プロセスに分けられ、その間をＲＰＣ（Ｒｅｍｏｔｅ　Ｐｒｏｃｅｄｕｒｅ　Ｃａｌｌ）によって通信されます。また、ディスプレイにはタッチパネルが装着されており、主に画面にタッチして使用します。（図４：エージェント用システムを参照）

2.3 クライアント用システム

クライアント用システムもエージェント用システムと同様に２つのプロセスに、分かれており、ディスプレイにはタッチパネルが装着されており、主に画面にタッチして使用します。

2.4 通訳者用システム

通訳者用システムは２台に分かれており、１台はディスプレイにタッチパネルが装着されており、マップやホテル予約シート、送られてくるテキストを表示するウインドウがあり、ボタンを押すだけのシステムにしています。（図２：通訳者用タスクシステムを参照）もう１台は音声合成用の文章を表示するウインドウや、自分でキーボードから打ち込んだテキストを相手に送る為のウインドウであり、主にキーボードを使用するようなシステムになっています。（図３：通訳者用音声合成システムを参照）

3 実験設定

この実験では、会話の目的をクライアントに説明し、まず基本的なコンピュータの操作に慣れる為に練習用のシステムに触れてもらいました。会話の目的は、クライアントが京都駅に置かれているこのシステム上で、国際会議が開かれる国際交流センターへの行き方をＡｇｅｎｔに尋ねること及び今回の国際会議用に用意してあるホテルの予約も行なうことと設定しました。エージェントには、あらかじめ実験前にシステムの操作方法に慣れてもらい全ての実験に参加してもらいました。通訳者にも、システムの操作方法に慣れてもらい全ての実験に参加してもらいました。

- 通訳者用システム

  通訳者用システム（通訳者＆合成音）。
  通訳者はもちろん、英語、日本語両方とも理解できる人。
  エージェントとクライアントの間に入り、橋渡し的役割を果たす。
  通訳者は、音声合成音を自由にクライアントへ送ることが出来ます。

- エージェント用システム

  国際会議事務局の置いてあるシステム（日本人で多少英語が理解できる人）。
  このシステムを使用して、クライアントの応対をする。

- クライアント用システム

  被験者用システム（英米人で、コンピュータに慣れている人と慣れていない人）。
  国際会議に出席するために今京都駅に着いた所である。
  現在地にこのシステムが置いてあり、国際会議場への案内やホテルの予約などを
  エージェントに質問する。

4 システム詳細

4.1 取り扱うメディア

音声、ビデオ画像、地図、ホテル予約シート、テキスト、音声合成音

1. 音声

音声構成は、クライアントのみスピーカとスタンドマイクで、通訳者とエージェントは
ヘッドホン付きマイクを使用している。
また、通訳者からクライアントへの音声は音声合成システムＣＨＡＴＲを使用している。
通訳者に聞こえる音声としては、クライアント，エージェントの声、それとクライアントのマイクに入る音声合成音。
クライアントに聞こえる音声としては、ＣＨＡＴＲのみ。（ただし、場所がエージェントと隣同士なので壁ごしに聞こえているかも。）
エージェントに聞こえる音声としては、通訳者の生の声とクライアントの声、それとクライアントのマイクに入る音声合成音である。
全ての音声は、ＤＡＴテープに収録されている。

2. ビデオ画像

ビデオ画像は、クライアント用システム，エージェント用システムの各マシン上に８ｍｍビデオカメラを置き、そこからＳケーブルを自分のマシンに、コンポジットケーブルを相手側のマシンにつなげる構成である。
将来、別々の遠隔地との通信実験をする場合は、ケーブルを伸ばすわけには、いかないので
イーサネットケーブル等を使用した通信プログラムを作成しなければならないだろう。
クライアント用システムには、エージェントの動画像と自分の静止画像。
なぜ、自分の画像が静止画像なのかは、ＮｅＸＴマシンにインストールされているビデオボードが
１つの動画像信号しか受け付けないからである。
もし、将来に他のマシンに移植する場合は２つの動画像が流れた方が好ましいと思う。
エージェント用システムには、クライアントの動画像である。
通訳者用システムには、ビデオ画像は何も表示しない。

3. 地図：

地図は、エージェント用システム、通訳者用システム、クライアント用システムの全てに表示されている。
地図の選択、クリアーはエージェント用システムのみが可能である。
地図へのペイントはエージェント用システム、クライアント用システムが可能である。
ペイントのカラーは、エージェント用システムがｒｅｄ、ｙｅｌｌｏｗ，ｓａｌｏｍｏｎの内のどれかで、選択することが出来る（デフォルトは、ｒｅｄ）。
クライアント用システムは緑に固定している。
通訳者用システムはペイント不可。ただし、エージェント，クライアントからのペイントをそれぞれ相手側に送ることが出来る。

4. ホテル予約シート：

ホテル予約シートは、エージェント用システム、通訳者用システム、クライアント用システムの全てに表示されている。
食事の選択のみ、エージェント用システムが可能である。
他のタグは、エージェント用システム、クライアント用システムとも可能である。
通訳者用システムは全て不可。

5. テキスト：

エージェント用システム、クライアント用システム、通訳者用システム全て可。ただし、エージェント用システムまたはクライアント用システムが送ったテキストは、全て通訳者用システムに送られる。それを、通訳者用システムは翻訳が必要でない限り、それぞれ相手側に送ることが出来る。

6. 音声合成音

音声合成音は、第２研究室で開発されたＣＨＡＴＲを使用しました。
通訳者用システムのみ送信可能である。

4.2 通信データプロトコル概要

4.2.1 データ構造

ソケット間通信を利用するにあたって通信プロトコルを固定しました。大きく分けて、SYSTEM 関係、TEXT 関係、MAP 関係、HOTEL 関係の 4 つに分けた。

JOB NUMBER	TO	FROM	ITEM NUMBER								

- JOB NUMBER

```
0 SYSTEM
1 TEXT
2 MAP
4 HOTEL
```

- TO FROM

```
0 ALL
1 AGENT
2 CLIENT
3 INTERPRETER TASK
4 INTERPRETER SYNTHISIS
5 SERVER
```

- ITEM NUMBER

JOB NUMBER が TEXT の場合は、そのテキストの Length 値が入る。

4.2.2 SYSTEM 関係メッセージ

```
SERVER : Server Process
ITP    : Interpreter Task Process
ISP    : Interpreter Synthisis Process
AP     : Agent Interface Process
CP     : Client Interface Process
```
と以下略す。

1. Server からのソケット接続後のリクエストメッセージ各 Process から Server へ自分の ID 番号を知らせる
   - SERVER - > ITP, ISP, AP, CP
   - ITP - > SERVER
   - ISP - > SERVER
   - AP - > SERVER
   - CP - > SERVER
2. サーバーからの Wait メッセージ
   - SERVER - > ITP, ISP, AP, CP

3. CP からの音声合成に関する応答メッセージ、ISP からの音声合成終了メッセージ (Interrupt or Repeat or Rephrase)
   - CP - > Server - > ITP, ISP, AP
   - ISP - > Server - > AP
4. AP からのメニュー選択メッセージ (Map or Hotel)
   - AP - > Server - > ITP, ISP, CP
5. CP からの終了メッセージ
   - CP - > Server - > ITP, ISP, AP
6. CP からのテレホンコール
   - CP - > Server - > ITP, ISP, AP
7. Server からのイニシャライズ終了メッセージ
   - SERVER - > ITP, ISP, AP, CP
8. AP からのテレホンコール終了メッセージ
   - AP - > Server - > ITP, ISP, CP
9. ISP からのマップイニシャライズウインドウのクローズメッセージとテキストイニシャライズウインドウのクローズメッセージ
   - ISP - > Server - > CP
10. ISP からのホテルイニシャライズウインドウのクローズメッセージ
    - ISP - > Server - > CP
11. ISP からのテキストイニシャライズウインドウのオープンメッセージ
    - ISP - > Server - > CP
12. ISP からの最初のエージェント画面ウインドウのオープンメッセージ
    - ISP - > Server - > CP
13. CP からの最初のエージェント画面ウインドウがオープンされた事を知らせるメッセージ、最初のエージェント画面ウインドウの音声合成が終了したことを知らせるメッセージ
    - CP - > Server - > ITP, ISP, AP
    - ISP - > Server - > CP
14. CP からの始めてマップウインドウがオープンされた事を知らせるメッセージ
    - CP - > Server - > ISP
15. CP からの始めてテキストウインドウがオープンされた事を知らせるメッセージ
    - CP - > Server - > ISP
16. CP からの始めてホテルウインドウがオープンされた事を知らせるメッセージ
    - CP - > Server - > ISP
17. ITP, ISP からの ASysTant アニメーションの向き変更を知らせるメッセージ
    - ITP - > Server - > CP

- ISP -> Server -> CP

18. ISP からの状態メッセージの送信
    - ISP -> Server -> ITP
19. ISP からの音声合成終了メッセージ
    - ISP -> Server -> CP
20. ITP からの状態メッセージの送信
    - ITP -> Server -> AP

4.2.3 TEXT 関係メッセージ

1. テキストの送信
    - AP -> Server -> ITP
    - CP -> Server -> ITP
    - ITP -> Server -> AP, CP
    - ISP -> Server -> AP, CP

4.2.4 MAP 関係メッセージ

1. AP からのマップファイル送信メッセージ
    - AP -> Server -> ITP, ISP, CP
2. AP, CP, ITP からのドラッグペイント位置の送信
    - AP -> Server -> ITP
    - CP -> Server -> ITP
    - ITP -> Server -> AP, CP
3. AP からのマップクリアーメッセージ
    - AP -> Server -> ITP, ISP, CP
4. AP, CP, ITP からのサークルペイント位置の送信
    - AP -> Server -> ITP
    - CP -> Server -> ITP
    - ITP -> Server -> AP, CP
5. AP, CP からのマップサイズ送信
    - AP -> Server -> ITP
    - CP -> Server -> ITP
6. AP からのマップウインドウのクローズメッセージ
    - AP -> Server -> ITP, ISP, CP

4.2.5 HOTEL 関係メッセージ

1. 名前の入力
   - AP -> Server -> ITP, ISP, CP
   - CP -> Server -> ITP, ISP, AP
2. 電話番号の入力
   - AP -> Server -> ITP, ISP, CP
   - CP -> Server -> ITP, ISP, AP
3. ホテル名の入力
   - AP -> Server -> ITP, ISP, CP
   - CP -> Server -> ITP, ISP, AP
4. チェックインの'月'の入力
   - AP -> Server -> ITP, ISP, CP
   - CP -> Server -> ITP, ISP, AP
5. チェックインの'日'の入力
   - AP -> Server -> ITP, ISP, CP
   - CP -> Server -> ITP, ISP, AP
6. チェックアウトの'月'の入力
   - AP -> Server -> ITP, ISP, CP
   - CP -> Server -> ITP, ISP, AP
7. チェックアウトの'日'の入力
   - AP -> Server -> ITP, ISP, CP
   - CP -> Server -> ITP, ISP, AP
8. チェックインの"時間"の入力
   - AP -> Server -> ITP, ISP, CP
   - CP -> Server -> ITP, ISP, AP
9. チェックアウトの"時間"の入力
   - AP -> Server -> ITP, ISP, CP
   - CP -> Server -> ITP, ISP, AP
10. 大人の人数の入力
    - AP -> Server -> ITP, ISP, CP
    - CP -> Server -> ITP, ISP, AP
11. 子供の人数の入力
    - AP -> Server -> ITP, ISP, CP
    - CP -> Server -> ITP, ISP, AP
12. シングルベッドの数の入力

- AP - > Server - > ITP, ISP, CP
- CP - > Server - > ITP, ISP, AP

13. ツインベッドの数の入力

- AP - > Server - > ITP, ISP, CP
- CP - > Server - > ITP, ISP, AP

14. 合計金額の入力

- AP - > Server - > ITP, ISP, CP
- CP - > Server - > ITP, ISP, AP

15. 連絡先の選択

- AP - > Server - > ITP, ISP, CP
- CP - > Server - > ITP, ISP, AP

16. 食事の種類の選択

- AP - > Server - > ITP, ISP, CP

17. ホテルウインドウのクローズ

- AP - > Server - > ITP, ISP, CP

5 ソケット接続 (イニシャライズ処理

まずServerを起動し、それから各Processを起動する。

1. あなたの番号は何ですか？

```
SERVER        -- >      ITP,   ISP,    AP,   CP
(0051)
```

2. 自分の番号をサーバーに送信する。

```
SERVER        < --      ITP,   ISP,    AP,   CP
                      (0531) (0541) (0521) (0511)
```

3. サーバーからのＷａｉｔメッセージ

```
SERVER
(0352)        -- >      ITP
(0452)        -- >      ISP
(0252)        -- >      AP
(0152)        -- >      CP
```

4. サーバーからのイニシャライズ終了メッセージ

```
SERVER        -- >      ITP,   ISP,    AP,   CP
(0059)
```

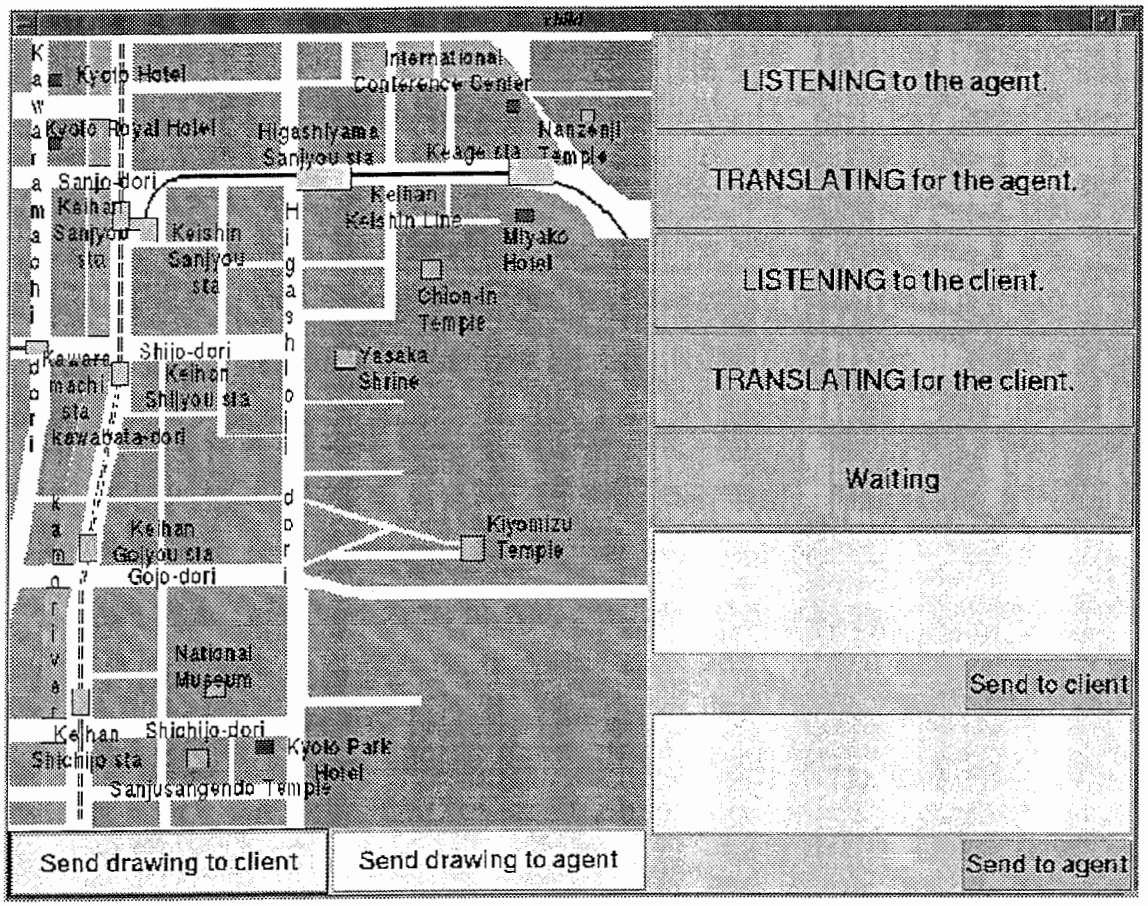

図 2: 通訳者用タスクシステム

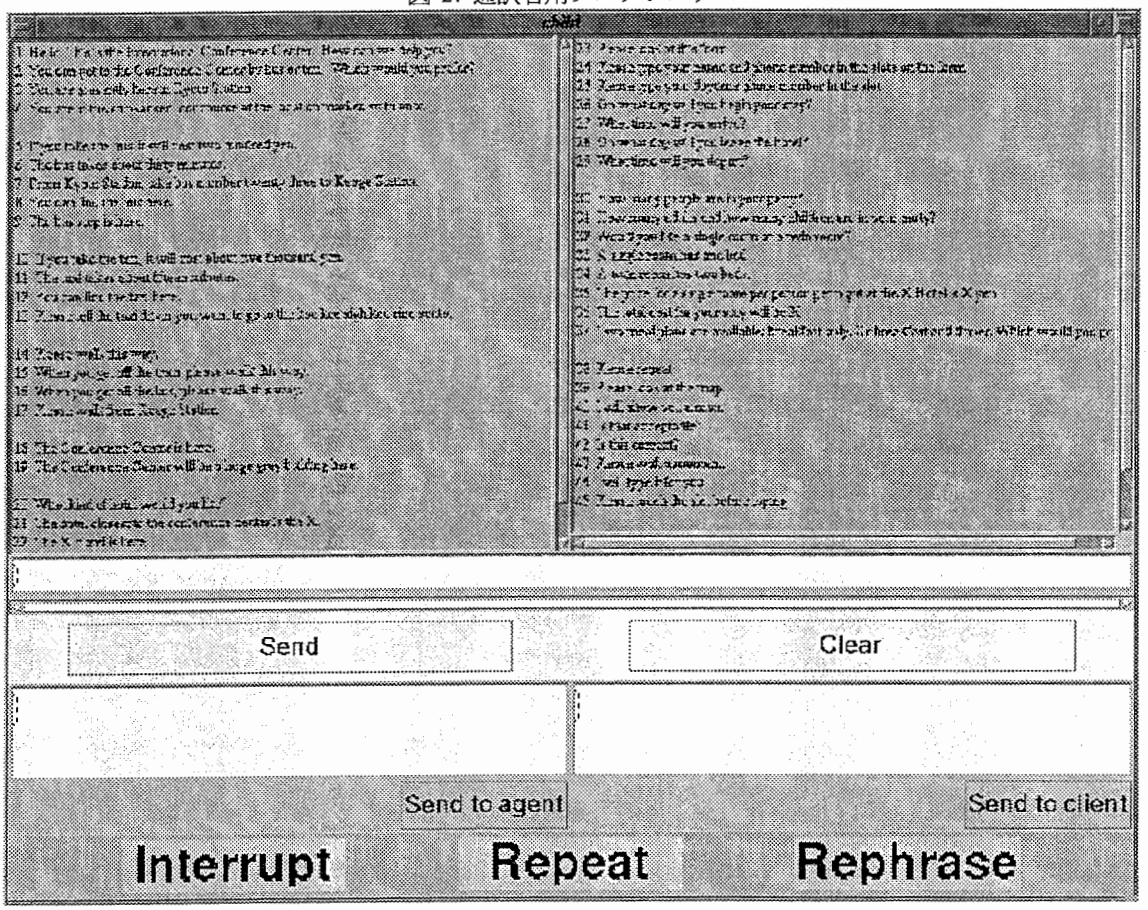

図 3: 通訳者用音声合成システム

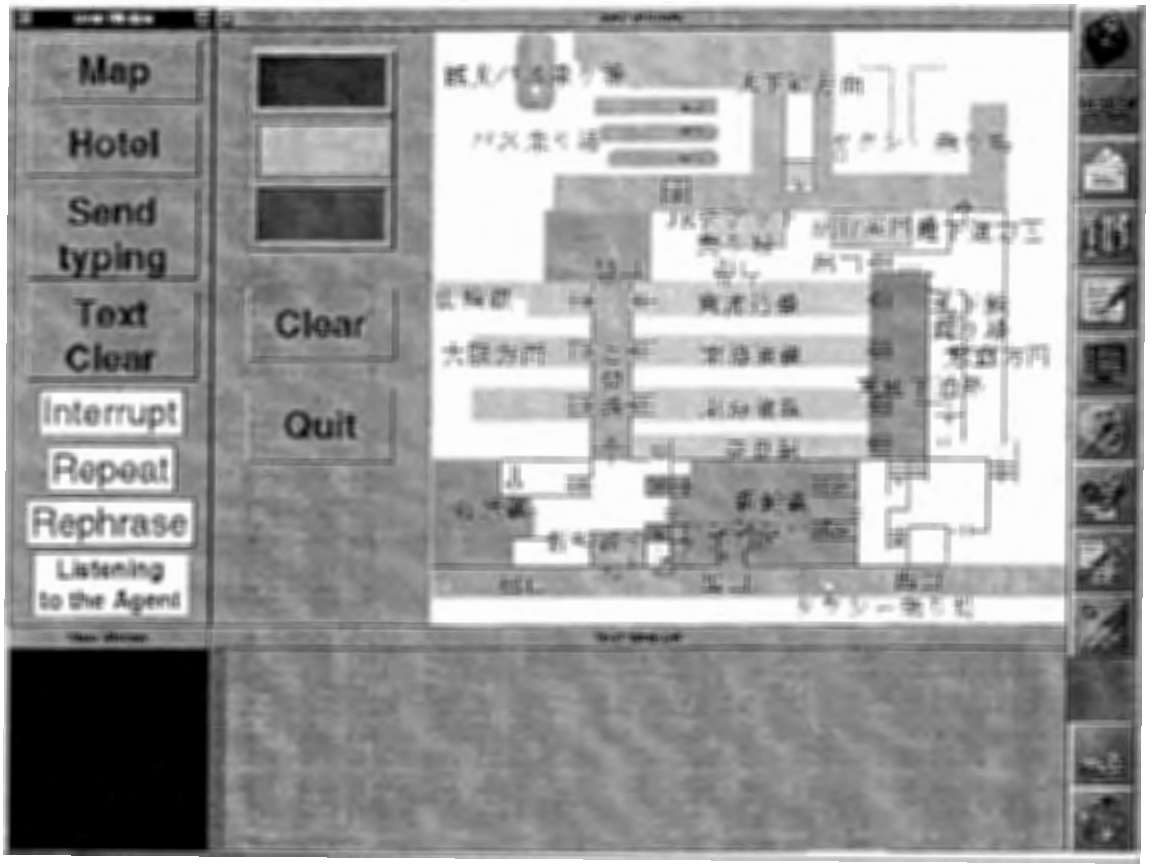

図 6: エージェント用システムスタート画面

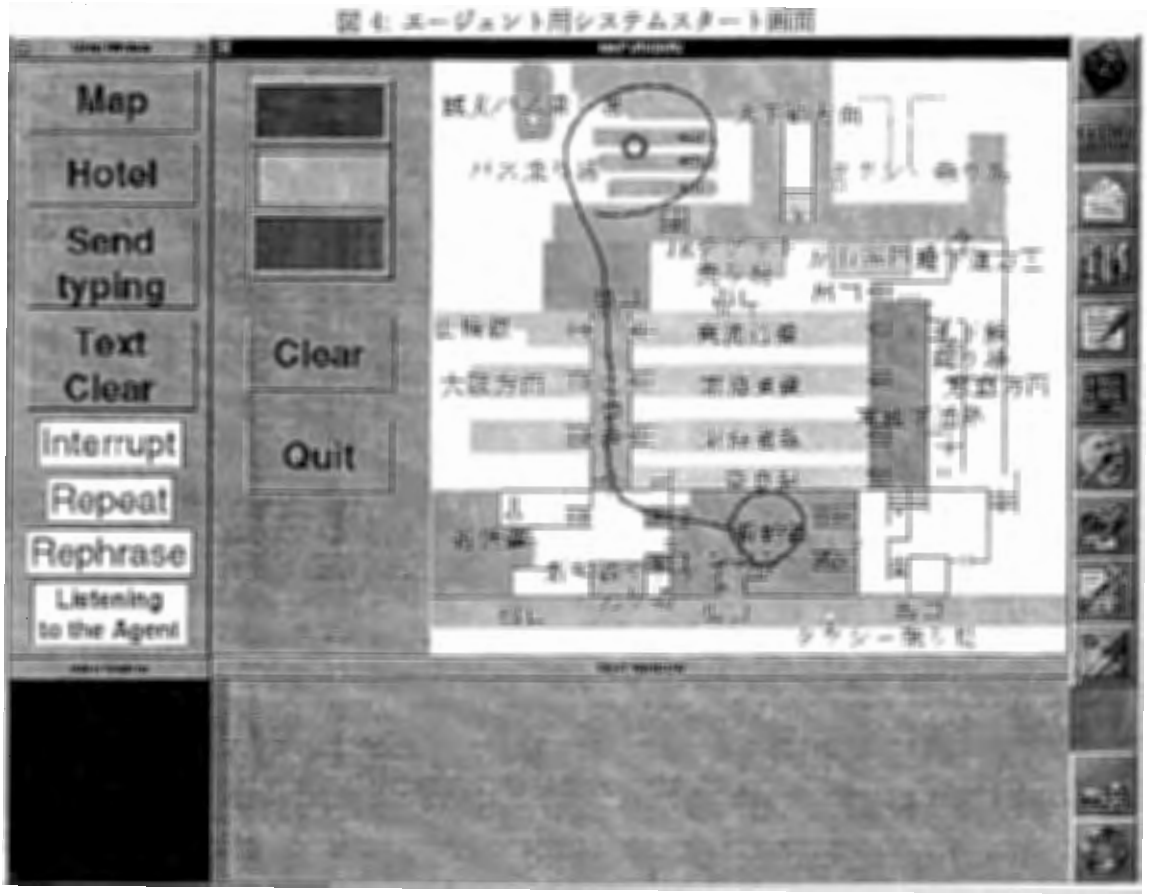

図 5: エージェント用システムマップ画面

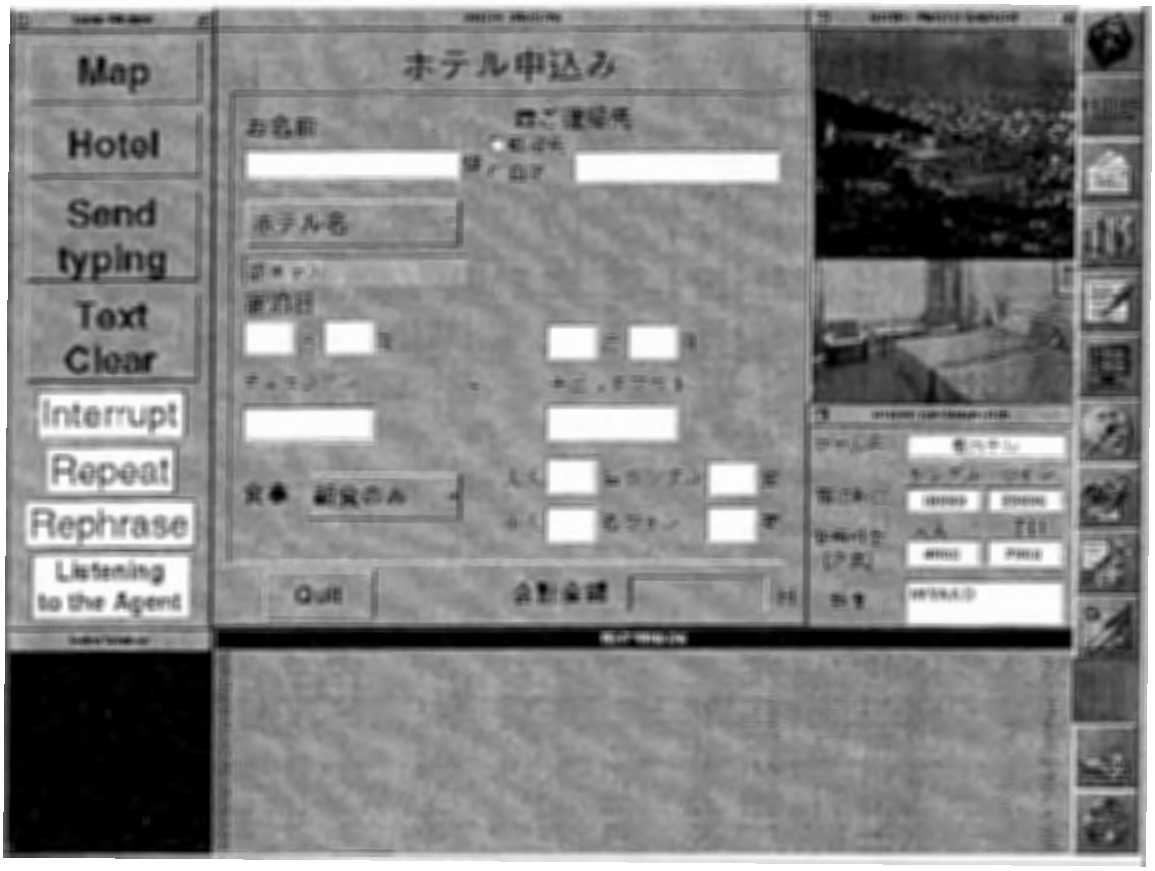

図 6: エージェント用システムホテル画面

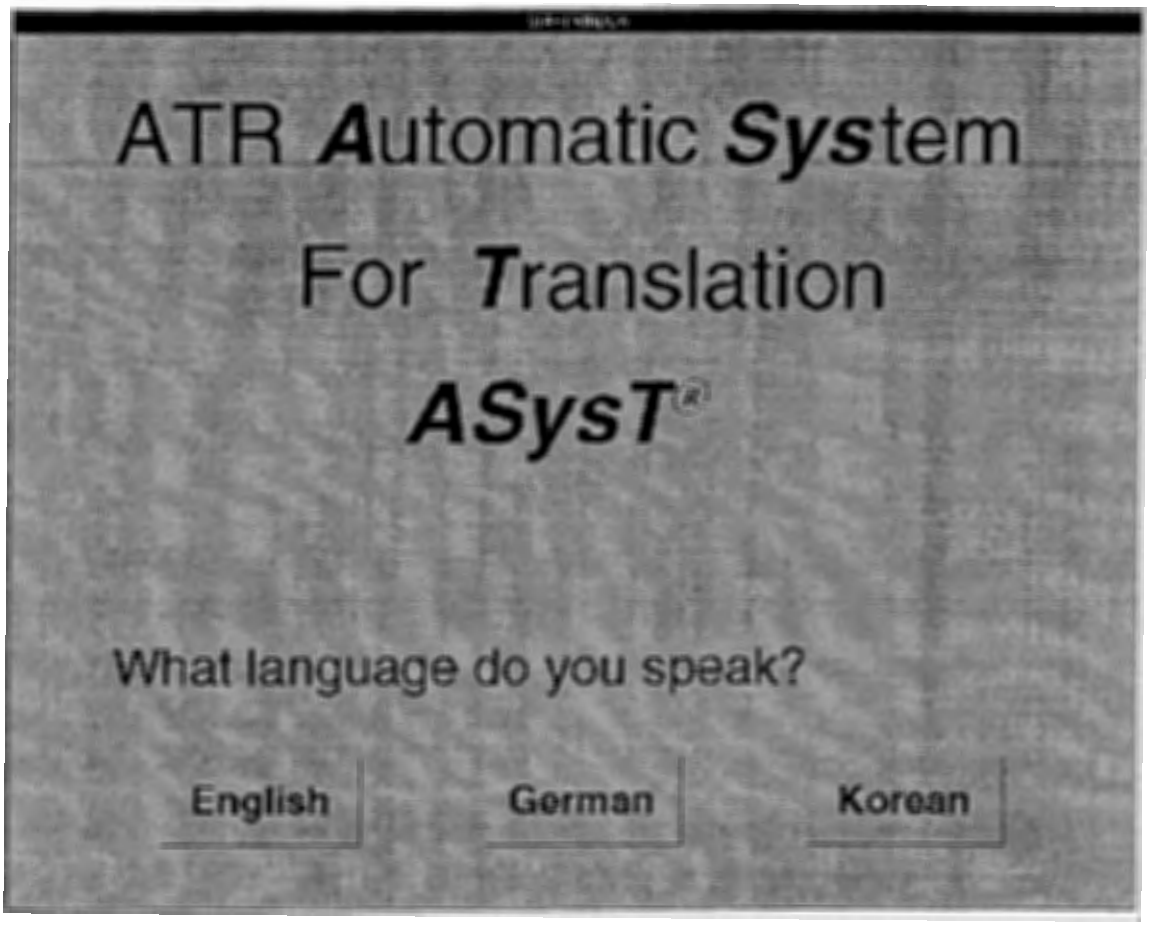

図 7: クライアント用システムスタート画面

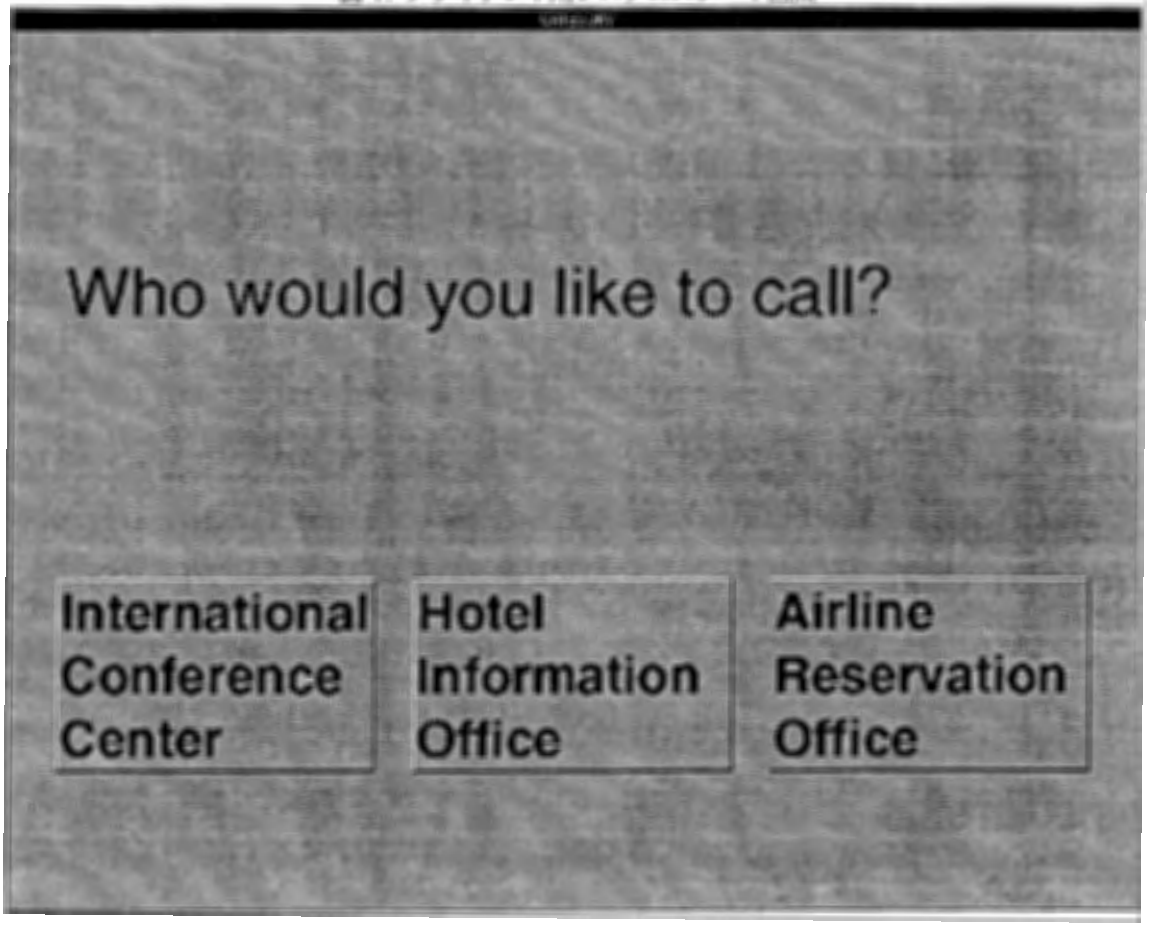

図 8: クライアント用システムタスク選択画面

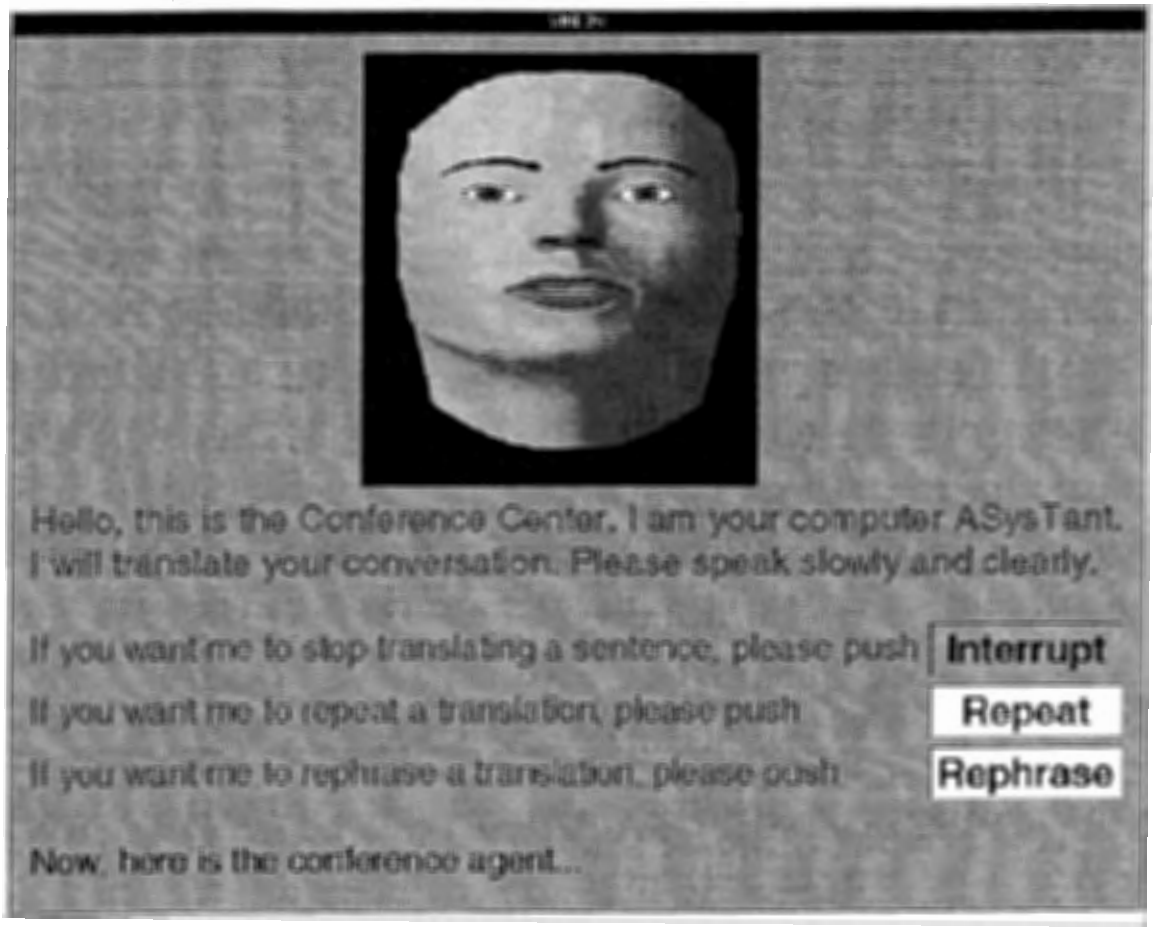

図 9: クライアント用システムイントロダクション画面

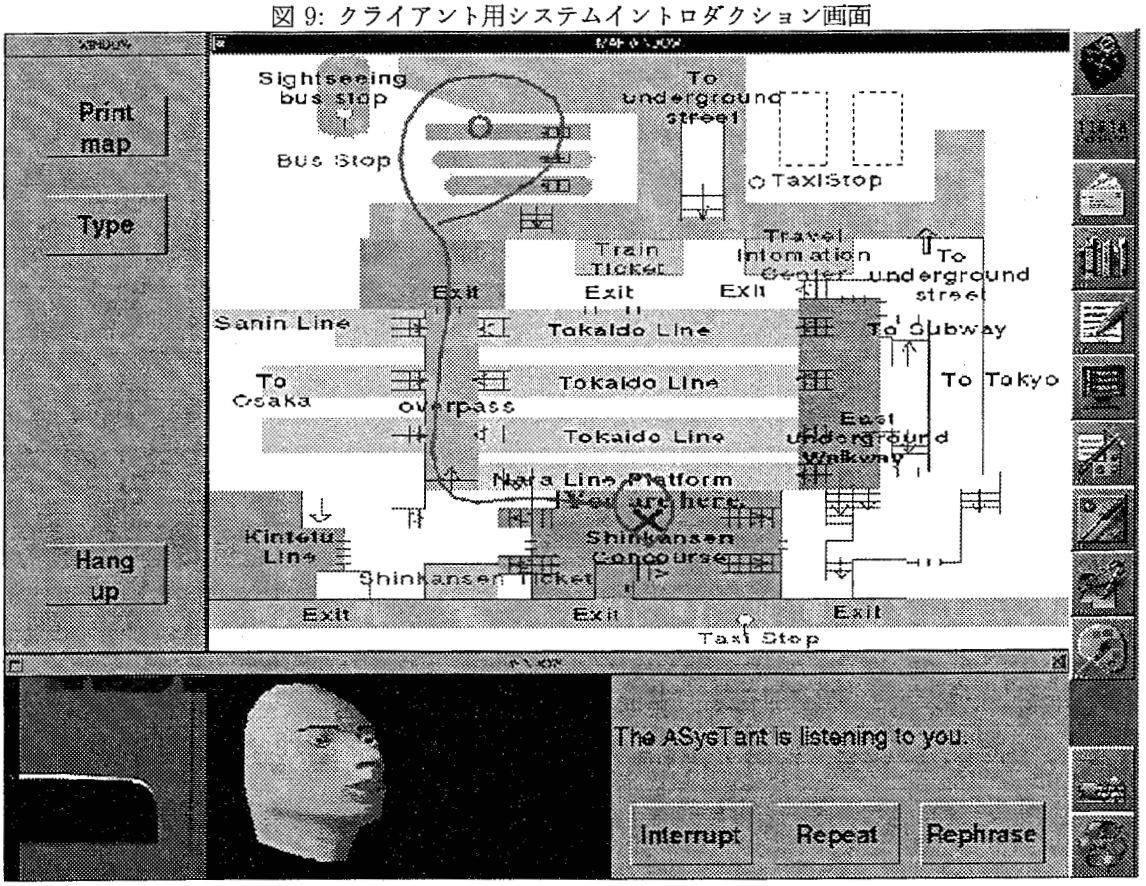

図 10: クライアント用システムマップ画面

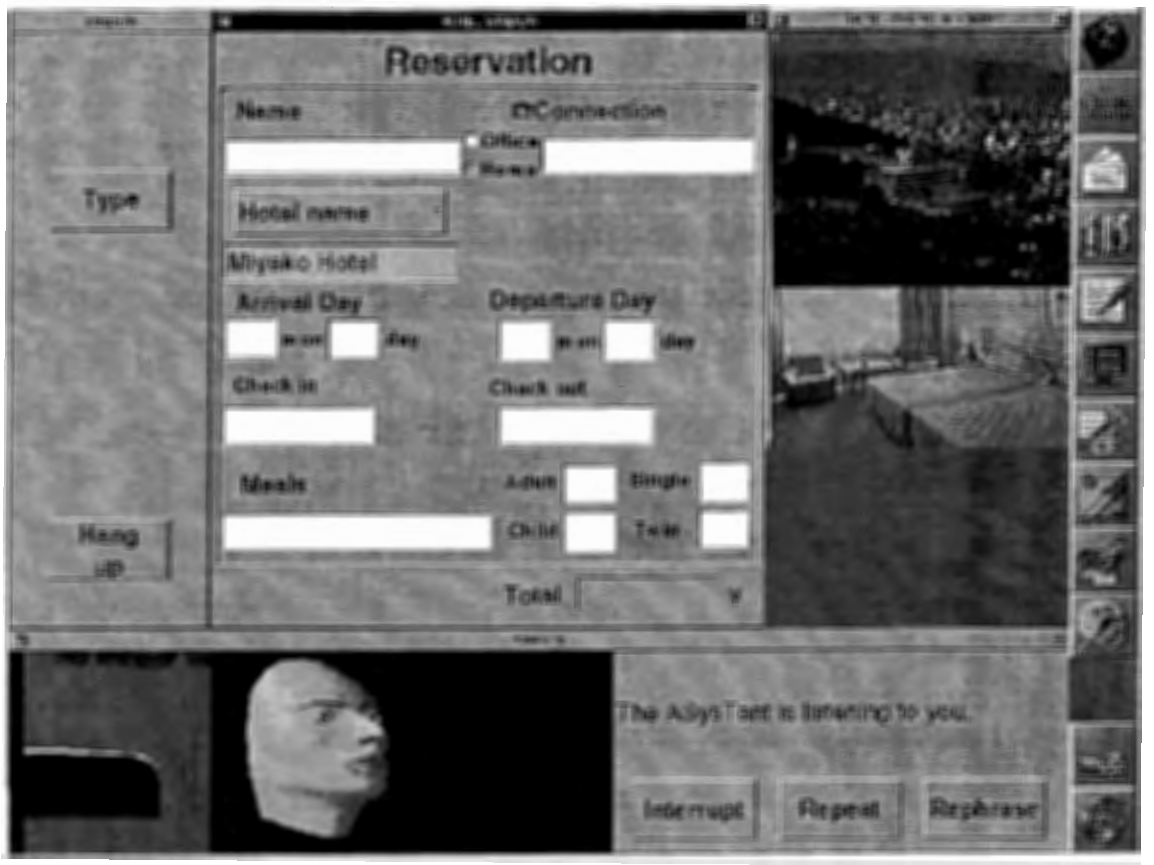

図 11: クライアント用システムホテル画面

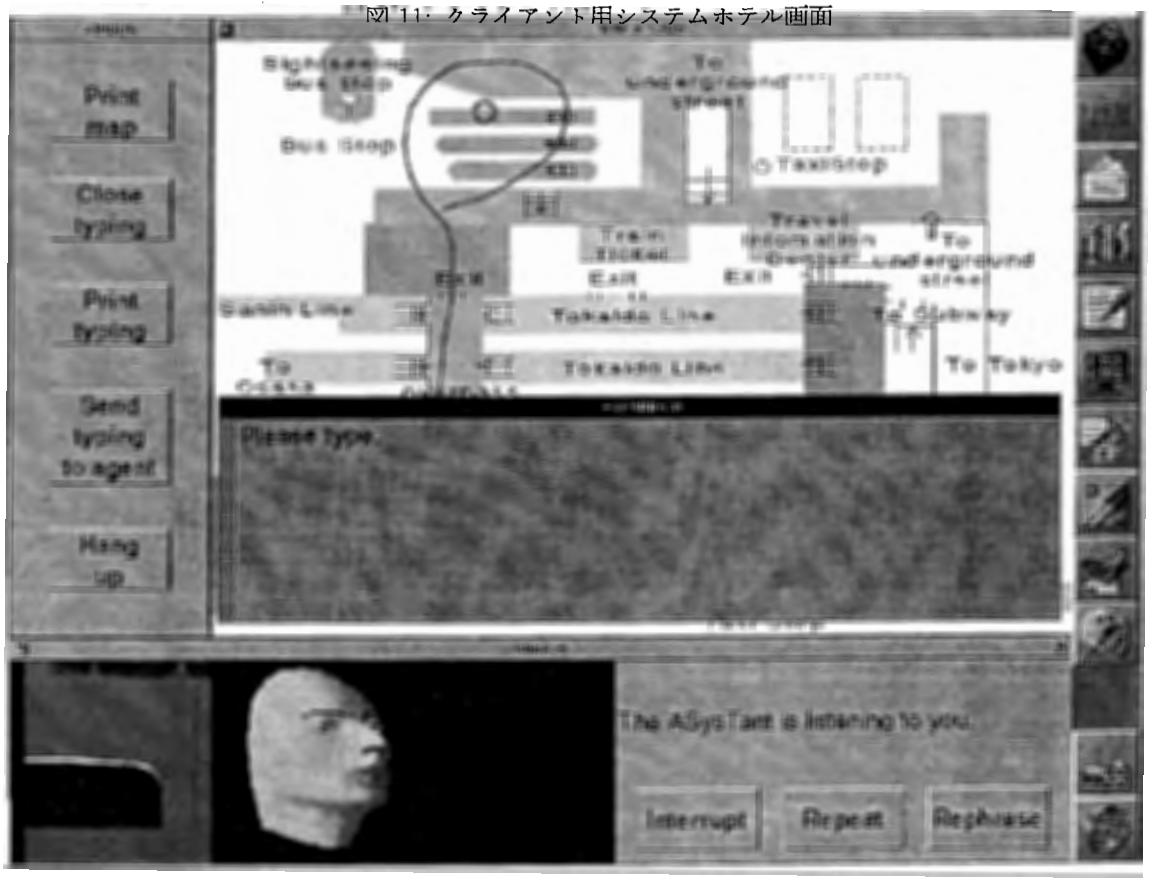

図 12: クライアント用システムテキスト画面

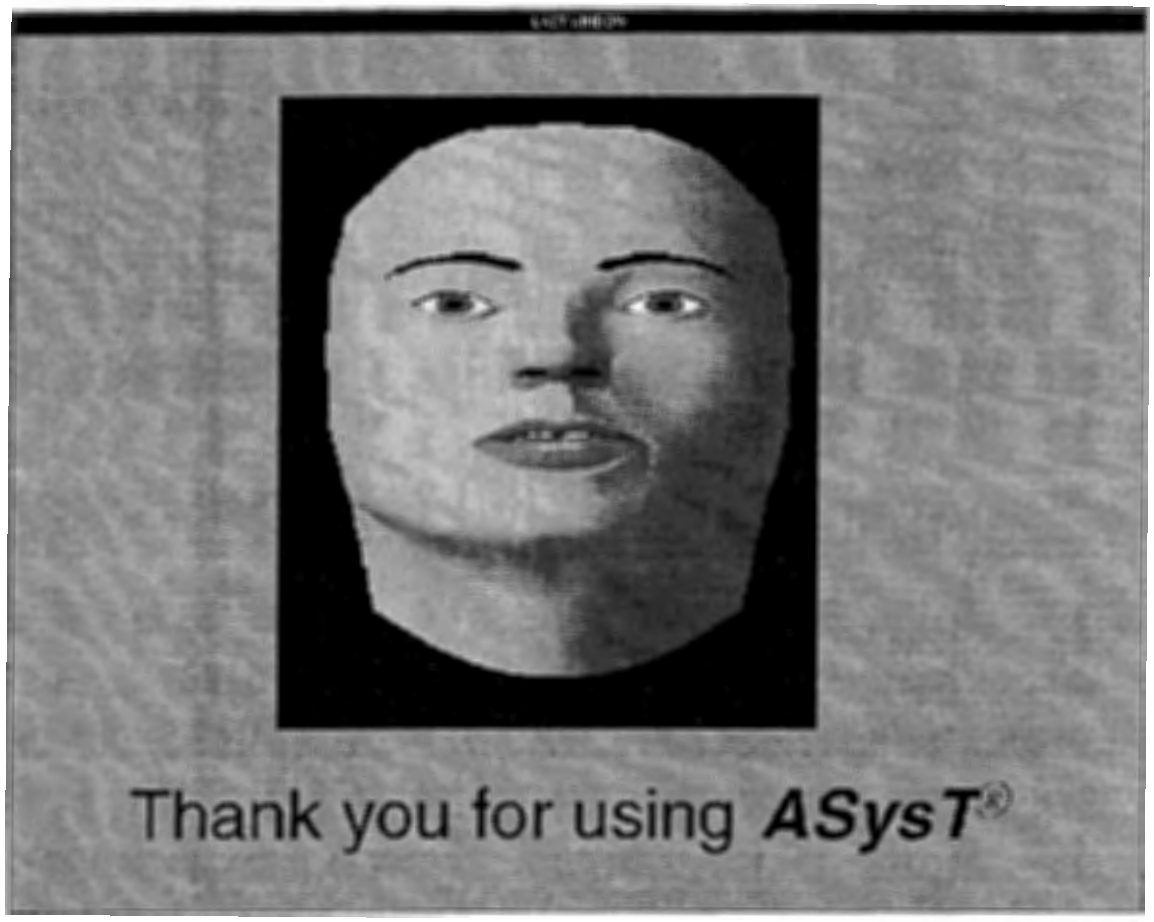

図 13: クライアント用システムラスト画面

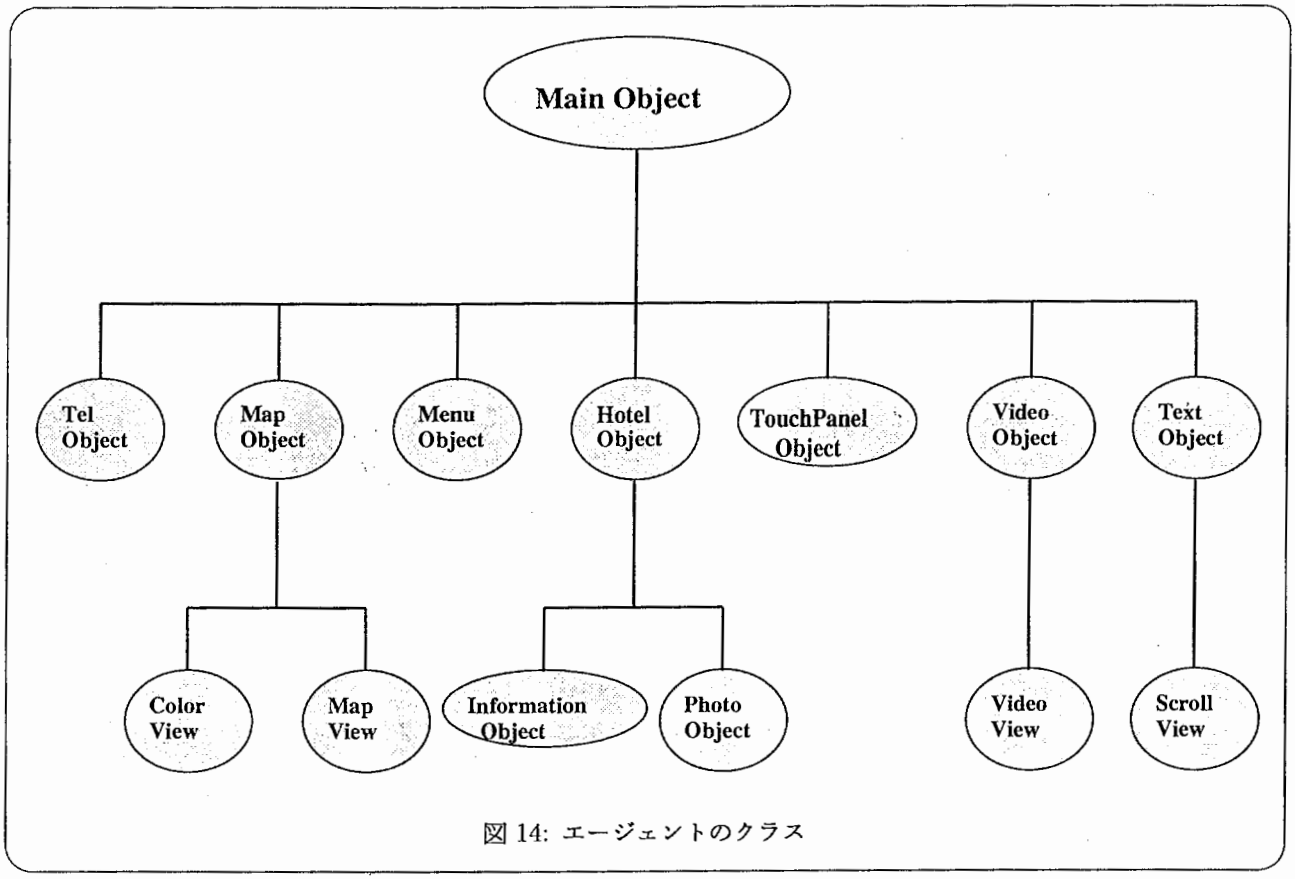

図 14: エージェントのクラス

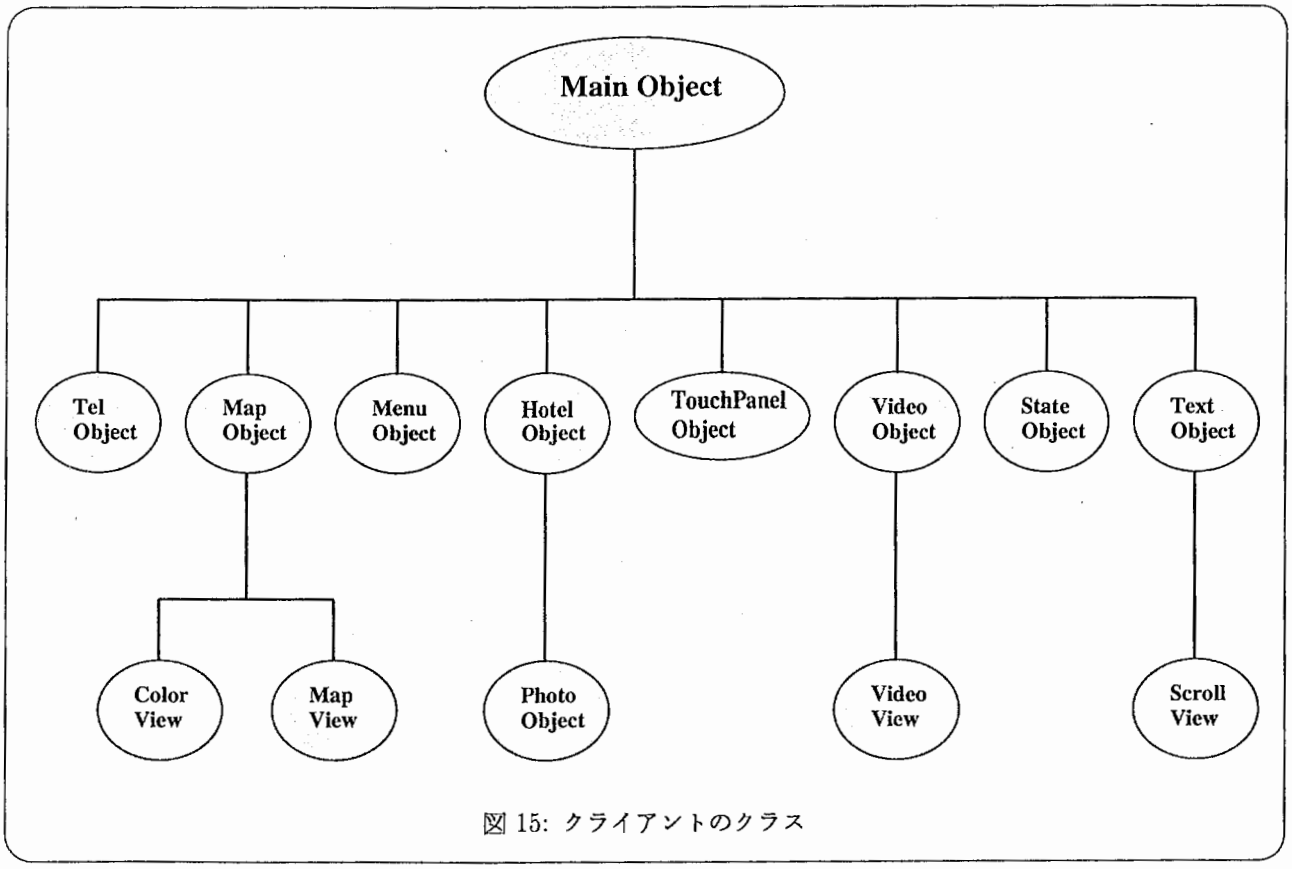

図 15: クライアントのクラス