音声言語情報処理 5－4
（1995. 2. 3）

Multimodal and Telephone-only Dialogues in English and Japanese

Kyung-ho Loken-Kim, Laurel Fais, Tsuyoshi Morimoto

ATR Interpreting Telecommunications Research Laboratories

We report on a comparison of English and Japanese speech in goal-directed conversation in both telephone-only and multimodal environments. Factors considered are: disfluency, distribution of syntactic sentence types, deictic expressions and distribution of utterance intention types. Responses to a subjective survey of subject's reactions to the communication modes are also discussed. Results may help us to construct models of speaker and hearer, as well as to understand the role of visual information in communication.

英語と日本語における マルチモーダル対話と電話対話の特徴分析

ローケン・キム キュンホ、 ローレル ファイス、森元 逞

エイ・ティ・アール 音声翻訳通信研究所

本稿では、ディスフルーエンシー、文の構文タイプ、指示表現インテンション・タイプ等に着目し、英語と日本語におけるマルチモーダル対話と電話対話の特徴を分析した結果を報告する。さらに、被験者のコミュニケーション・モードに対する感想に関しても言及する。本研究の結果は、話者モデルの構築や、通信における視覚情報の役割の理解に貢献できると考える。

1. INTRODUCTION

The work reported here is part of an investigation of linguistic and communicative behavior of speakers of English and Japanese via telephone and multimedia environment in EMMI (Environment for MultiModal Interaction) [1,2]. There is much previous work to suggest that speakers do, in fact, accommodate their speech to the communication environment in which they are interacting (see [3] for a survey). Oviatt et al. [4] report that linguistic variability in people's speech and writing can be reduced with the careful selection of input modality and presentation format. More specifically, Oviatt [5] found an approximately 45% lower disfluency rate (per 100 words) in form-based human-computer interaction than in unconstrained human-computer interaction. This confirms the natural assumption that the nature of the communication environment influences the carefulness and planning with which humans conduct their conversations.

The work reported on here is an attempt to discover if the kind of differences in fluency apparent in Oviatt's work would be found as well in a comparison of situations where communication was much less restricted, but still possibly influenced by the mode through which it was taking place.

We were interested, as well, in investigating other possible linguistic differences characteristic of speech in each mode: the distribution of syntactic sentence types and the use of deictic expressions. We also examined the paralinguistic measures of frequency of use of speaker intention types and a number of parameters of ease and comfort of use for each mode. And, of course, for all these various measures, an important question to ask was: are these modal differences the same across English and Japanese? It would be optimal, of course, to be able to design a multimodal natural language processing system which did not require modification for particular languages. Part of the aim of this investigation was to begin to see how far this would be possible.

2. METHOD

Sixteen subjects, eight native speakers of American English and eight native speakers of Japanese took part in the experiment. They were told to imagine that they had arrived in Kyoto Station, having never been there before, and that they had to find their way to a conference described on a "brochure" they had been given. Their sole means of acquiring this information was by talking to the "conference agent" at the "conference office." None of the subjects knew the agent, nor were they at all familiar with EMMI. The subjects were told that they were to play the part of client twice, once in a telephone situation and once via the multimodal interface. (The full instructions appear in [6].) For each language group, four subjects participated in the telephone situation first; four used the multimodal set-up first. The agents in all trials were trained native speakers; one Japanese agent participated in all the Japanese trials; one American agent participated in all English trials.

In the telephone condition, subjects spoke into standard telephones, wearing a Sennheiser HMD 410 headset with microphone (one ear piece was turned up to allow for the telephone handset). In the MM environment, subjects spoke into the same headset-mounted microphone as in the telephone condition, but listened to the agent through the attached headphones. They sat in front of a NeXT computer monitor, with keyboard and mouse. On the screen appeared a video image of the agent with whom they were talking, a field for typing in written input, and an area in which several different maps could be displayed by the agent. Subjects could draw on the map by dragging with the mouse, could type on the keyboard, or could use speech to communicate with the agent (who had the same options for communicating with the subjects). Clients' and agents' drawings were differentiated by color. Subjects were also allowed to practice with the drawing and typing capabilities of EMMI until they felt comfortable.

Acoustic speech data was recorded on digital audio tapes using a SONY DAT deck, DTC-77ES. The acoustic tapes of the experiment sessions were transcribed, including notations for false starts; filled pauses such as "ah" and "uhum;" non-speech noises such as deep breaths or lip smacks; and simultaneous speech. The transcriptions were checked twice, all by independent transcribers. At the completion of each trial, subjects were interviewed to gather information concerning their reactions to each mode.

The full description of EMMI for this experiment can be found in [1], [2], and the full set of transcriptions is given in [6].

3. MEASURES

The linguistic differences between multimodal dialogues and telephone-only dialogues were measured for both English and Japanese by counting the number of: spoken disfluencies, including interjections (e.g., "ah", "eh", etc.), false starts and hesitations; syntactic types for sentences; and deictic expressions used. Paralinguistic differences in distribution of utterance intention types [7] were also examined.

4. RESULTS

1. Linguistic Differences

Disfluency. The standard measure for disfluency in English is the number of false starts made and and the number of filled pauses used per 100 words. However, the notion "word" is not as well established for Japanese at it is for English; thus defining a rate of disfluency in this way is undesirable for Japanese. For this reason, we measured disfluencies per turn rather than per 100 words. While this does not give us a measure comparable to the per word measure, it does give us a consistent measure for use in both languages.

Results for disfluency rates are shown in Figure 1 (results for the agent are averaged over all the conversations in which each agent took part; results for the client are averaged over all clients, for each mode and language). In English, the clients' disfluency rates were overall lower than those of the agent, and remained fairly constant across both modes. However, the agent showed somewhat lower disfluency rates in the telephone than in the MM mode for English.

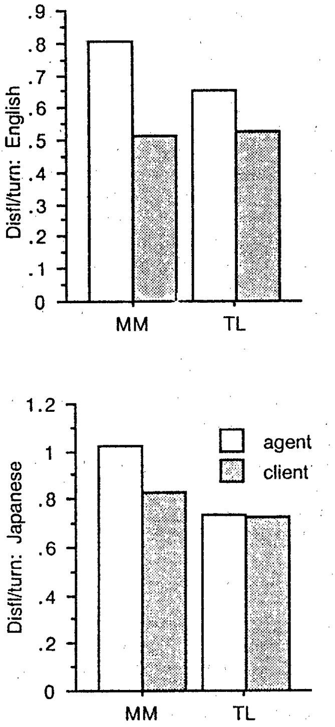

Figure 1. Disfluency per turn for English (top) and Japanese (bottom)

Results for Japanese were slightly different. As for English, client disfluency was overall lower than agent disfluency in the MM mode, but the difference between client and agent was negligible in the telephone mode. And, again similar to the English case, the agent's disfluency was greater in the MM mode. The major difference in the Japanese case seems to be the similarity in rate between client and agent in the telephone mode.

Sentence syntactic types. All dialogues were labelled by hand for syntactic classifications. The categories used were "declarative," "interrogative," "imperative," and "idiom." Because imperatives occurred only extremely rarely, they are not included in Figure 2. "Idiom" denotes expressions which are "none of the above;" that is, they are short expressions such as "thank you" or "OK," as well as utterances which are not full declaratives, such as "Kyoto subway station."

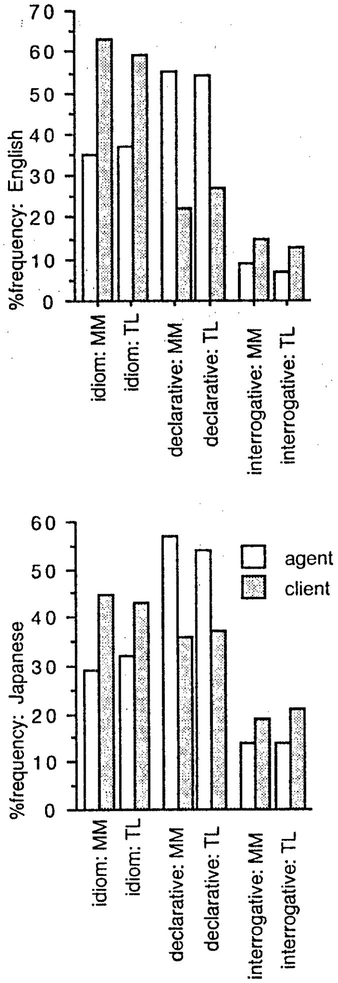

Figure 2. Per cent frequency of syntactic sentence types for
Englis (top) and Japanese (bottom)

The analysis of syntactic types show similar results for both English and Japanese. The agent, acting as an information provider, used declarative sentences more than half of the time in both telephone and and MM modes, while the client, in responding to the agent, most often uttered short idiomatic sentences to respond to the receipt of information. The client also asked more questions than the agents, as would be expected from the person in the role of information requester.

There was only slight variation between Japanese and English, mostly in the form of a higher percentage of declarative utterances and proportionately lower percentage of idiomatic utterances in the Japanese client case. This is an artefact of the syntactic structure of the forms used in Japanese to respond; in the polite case necessitated by this conversational setting, these are more often in full sentence form than those in English.

Deictic expressions. The frequency of use of the various deictic expressions in Japanese and English was counted (Figure 3). For English, these expressions are the near deictic locative and (demonstrative) pronominal forms, "here" and "this," respectively, and the far locative and (demonstrative) pronominal forms "there" and "that." Japanese has similar sets of near and far deictic expressions. Near forms are prefixed by ko: kore, "this" (pronominal); kono, "this" (demonstrative), and kochira, "here" or "I." Far forms are prefixed by so: sore, "that" (pronominal); sono, "that" (demonstrative), and sochira, "there" or "you."

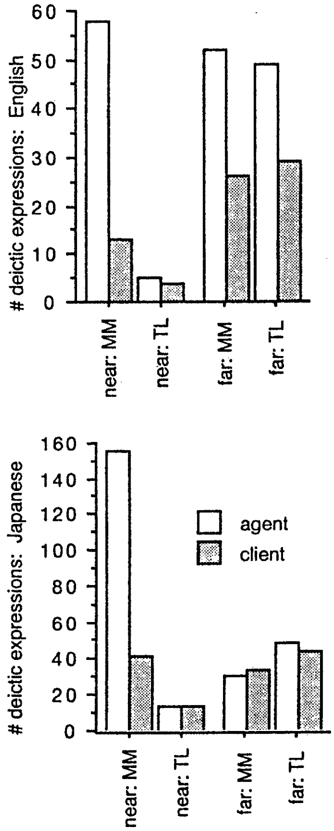

Figure 3. Number of deitic expressions used in English (top) and Japanese (botton.)

In Japanese, deictic expressions were used in approximately the same way for both modes by both client and agent, except in the case of the use of the near deixis forms by the agent. This exceeded the use of these forms by the client in this case. This is consistent with the agent's role as information giver.

The English situation was slightly more complicated. For the use of near deictic expressions, it was similar to Japanese, with the agent again using a much higher number of near forms in the MM mode. However, the agent also used a much greater number of far forms in both modes than did the client.

It seems to be the case that in both English and Japanese, near deixis is not used frequently in telephone mode; speakers using pronouns are often using them to refer to elements in the speech of others (or to refer to elements in their own speech, from the point of view of others), and thus there is little need for near forms in telephone speech. With the introduction of the shared graphics and the (near) visual objects in the MM mode, speakers change to using near forms more frequently, especially when the speaker is in an information-giving capacity.

On the other hand, the differences in the use of far forms may be language-based. All the clients tended to use far forms more than they used near forms in the telephone mode. But English agents use these forms much more than English clients. This seems to be consistent with the information-giving role of the agent. The lower frequency of use for far deictic expressions by Japanese agents may again be an artefact of the structural composition of Japanese utterances.

2. Paralinguistic Differences

Intention types. We used the labelling system for utterance intention types developed at ATR [8] . The labels which appear in Figures 4 and 5 are fairly self-explanatory, with the possible exception of "expressive," which is used to denote utterances such as "wow," or "oops." The labelling system itself consists of 27 labels with explicit conditions for their application; only those labels which accounted for at least 3% of the total utterance types for either language are included in Figures 4 and 5. "Instruct" is defined for English, but not for Japanese in this labelling system; this is the reason why Japanese speakers are shown using no "instruct" utterances.

All other cases which show 0% represent possibilities which are not realized in these dialogues, or which appear so rarely as to make their frequency virtually 0%.

Differences in the distribution of utterance intention types seemed to be linked primarily to language; there were no differences that could be reliably attributed to mode, which also held across both languages.

A comparison of the distributions of "inform" and "acknowledge" support the conclusions drawn from the analysis of syntactic sentence type above: agents, as information providers, used more "inform" type utterances, and clients, as information receivers, used more "acknowledgement" type utterances (often of the syntactic type "idiom," such as "OK" or "hai"). There were language-specific modal differences; English speaking clients "informed" much more often in telephone than in MM mode, while Japanese speaking agents "informed" more in the MM mode.

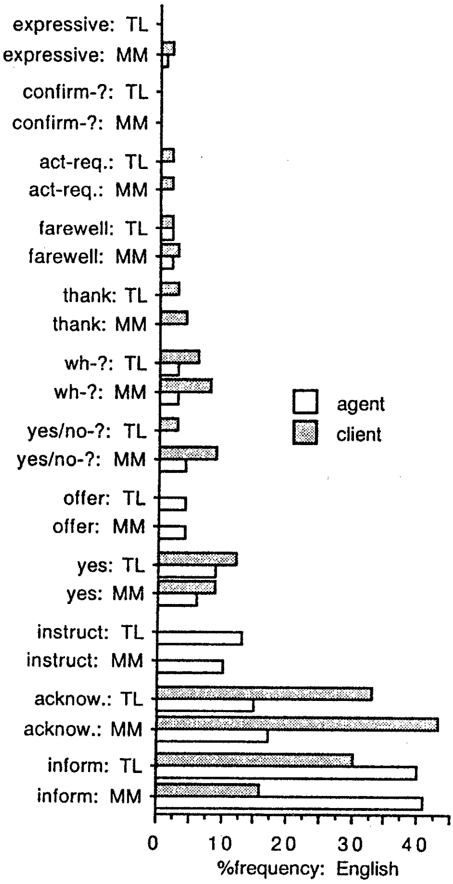

Figure 4. Per cent frequency of utterance in intention types for English

Both English- and Japanese-speaking clients asked more questions of both the yes/no form and the wh-form, and "thanked" more than agents, as is consistent with their roles as information receivers. It is also consistent with speakers' roles that agents are the only speakers to make "offers" in both languages; the larger number made by English-speaking agents probably reflects a cultural difference. Similarly the frequent use of "action-requests" by Japanese-speaking agents as compared to none by English-speaking agents is probably due to the fact that, in the Japanese dialogues, "please wait a minute" is much more frequent and was labelled as an "action-request." Likewise, the greater use of "confirmation-questions" by Japanese speakers is a result of the frequent use of the sentence-final particle "ne" used for this purpose. The English equivalent "right?" is used much less frequently.

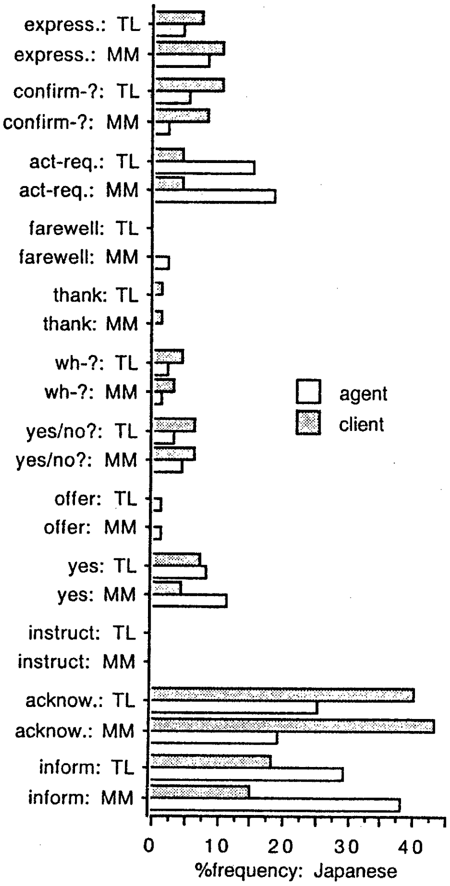

Figure 5. Per cent frequency of utterance intention types for Japanese

Ease of use. There was some variation in the impressions of both English-speaking and Japanese-speaking subjects regarding how easy it was to use EMMI. Most felt it was "simple," while a few felt it required "some effort." No one rated either mode as "difficult." This result may have been due in part to the fact that all the subjects were familiar with computers. Some subjects apparently felt that the manipulation required to enable typing was somewhat confusing or complicated.

Video image. Surprisingly, only a few subjects commented on the video image; they seem to appreciate having the video image of the agent available, but were divided as to how useful or necessary it actually was.

Map. Virtually across the board, subjects reported finding the map very useful, much more so than the auditory channel alone (telephone mode). Although they occasionally experienced some difficulties in using the map, and requested the option of getting a copy of the map, subjects still felt that this visual channel was extremely worthwhile.

Hard copy. In both the telephone and MM modes, many subjects took some notes on scratch paper while they were engaged in the dialogue. After the experiment, some subjects remarked that they wished they had taken more notes or had a hard copy of the map. The suggestion to include printing capability to the system was a frequent one.

Cursor position, mouse, and keyboard. Several subjects had difficulty finding their own as well as the other party's cursors, especially when the cursors were inactive on the map. This is crucial because the initial cursor position on the map can be used as "You are here" sign, since many clients also had a hard time finding their own position on the map.

Subjects were also uncertain about drawing on the map. When they drew a line on the map using the mouse, they were not sure whether the line appeared on the other party's map, giving rise to confirmation messages such as "Do you see this line?".

Most of the subjects felt that the keyboard was not very useful. This may have been partly due to the difficulties involved in accessing the keyboard, but it may also have been due to the fact that the task did not require the use of the keyboard, and subjects simply took the path of least resistance and did not use it.

5. DISCUSSION

The results of the questionnaire seem to indicate that subjects did enjoy using the system and found especially the map to be quite useful. This would indicate that there is some sense in which communication is more effective in the MM mode.

On the other hand, the objective results reveal a complex picture of the effects of mode on spontaneous conversation in these two languages. It is apparent that the MM environment in and of itself is not sufficient to reduce users' disfluency rates, unlike the form-based condition in Oviatt [5]. Subjects showed similar levels of disfluency for both the telephone and MM modes, with the agents' disfluency rates being somewhat higher for the MM mode. (This result is corroborated for English using a disfluency-per-100-words criterion [9]).

As would be expected, the results of the analyses of both utterance intention and syntactic structure support the view of the clients as information receivers, responding with idiomatic phrases (or in the case of Japanese, with idiomatic phrases and short declarative sentences). In the English case, this suggests that recognition of key phrases and fragments, what have been labelled here as "idioms," will be essential since they make up such a large portion of the clients' speech; this is somewhat less the case for Japanese.

The analysis of deictic expressions shows a marked trend toward a much higher frequency of use for near deictic expressions in the MM mode. This is consistent with the finding that, in the MM mode, the visual channel takes over some of the burden of communication that is carried by the speech channel in the telephone mode [9] . The use of near deictic expressions marks instances in which the visual channel rather than the speech channel is carrying the information in the exchange. The differences in the distribution of deictic expressions, therefore, seems to be correlated with differences attributable to the communicative mode of the conversation.

How can we incorporate these findings into the design of a spoken language interpretation system [10] ? The following results suggest the potential for introducing multimodality to a spoken language interpretation system: 1) The clear distinction of the speakers' roles in the multimodal dialogues can help us to define speaker models. 2) The lower disfluency rates for clients when compared to those for agents are encouraging; while they were not, as hoped, lowered significantly in the MM mode, they at least were not elevated. (Of course, the agent can be trained to talk to a language processing system, which is, of course, a much easier task than training the naive general public). 3) The parallel use of modalities in the MM mode (mainly marking and speaking) increased the number of referent expressions, opening up the possibility of using the information provided by one modality to resolve the ambiguity occurring in the information provided by another modality.

There are areas, nevertheless, which could be problematic. For example, synchronizing multiple sources of information in a bilingual context requires a deep understanding of the information conveyed in the source language in order to translate it to the target language in a well coordinated manner. Without such information understanding and synchronization, the bilingual multimodal interactions will be at best confusing and at worst incomprehensible.

Acknowledgement

The authors would like to express sincere gratitude to Dr. Yamazaki, the president of ATR Interpreting Telecommunications Research Laboratories, for the support and cheerful encouragement which made this study possible. We also would like to thank Mr. Kurihara for his efforts in realizing EMMI.

Bibliography

1) K. H. Loken-Kim, F. Yato, K. Kurihara, L. Fais, and R. Furukawa. EMMI-ATR environment for multi-modal interactions. ATR Technical Report TR-IT-0018.

Kyoto, Japan: ATR Interpreting Telecommunications Research Laboratories. 1993.

2) K. H. Loken-Kim, F. Yato, and T. Morimoto. A simulation environment for multi-modal interpreting telecommunications. IPSJ, AV Multiple Information Processing Workshop 4-1, 3/18, 1994. (in Japanese)

3) H. Giles, A. Mulac, J. J. Bradac, and P. Johnson. Speech accommodation theory: The first decade and beyond. In Communication Yearbook 10, ed. by M. L. McLaughlin. Sage Publications. Newbury Park. 1987.

4) S. Oviatt, P. Cohen, M. Wang and J. Gaston. A simulation-based research strategy for designing complex NL systems. ARPA Workshop on Human Language Technology, March, 1993.

5) S. L. Oviatt. Predicting spoken disfluencies during human-computer interaction. Proc. CHI '94, Boston. 1994.

6) K. H. Loken-Kim, F. Yato, L. Fais, Kazuhiko Kurihara, Ryo Furukawa, and Yoshihiro Kitagawa. Transcription of spontaneous speech collected using a multi-modal simulator--EMMI, in a direction-finding task (Japanese-Japanese; English-English). ATR Technical Report TR-IT-0029. Kyoto, Japan: ATR Interpreting Telecommunications Research Laboratories. 1993.

7) M. Nagata, M. Suzuki, and S. Tsukawaki. First steps toward annotating illocutionary force types to a bilingual dialogue corpus. ATR Technical Report TR-I-0298. Kyoto, Japan: ATR Interpreting Telecommunications Research Laboratories. 1993.

8) M. Seligman, L. Fais, and M. Tomokiyo. A bilingual set of communicative act labels for spontaneous dialogues. ATR Technical Report TR-I-0081. Kyoto, Japan: ATR Interpreting Telecommunications Research Laboratories. 1994.

9) L. Fais. Effects of communicative mode on spontaneous English speech. Technical Report of the Institute of Electronics, Information and Communication Engineers, NLC94-22(1994-10). 1994.

10) F. Yato, T. Morimoto, Y. Yamazaki and A. Kurematsu. Important issues for automatic interpreting telephone technologies. Proc. ISSD-93, Nov. 1993.