TR-IT-0059

Effects of Mode on Spontaneous English Speech in EMMI

Laurel Fais and Kyung-ho Loken-Kim

June 1994

Results of a pilot study conducted in the ATR EMMI with English speakers are discussed. Significant patterning with respect to mode and order of disfluency rates, words used per goal reached, overall words used, and overall goals reached are reported. The implications of these results for multimodal interface design are discussed in conjunction with results from a post-experiment questionaire which elicited from the subjects judgments as to how enjoyable each mode was; how easy each mode was to use; and how useful each mode was.

Contents

Introduction

Current technology has made available a wide variety of devices with which communication between humans can be enhanced: video imaging, computer keyboard-based systems, touch screens, pen-based technologies, and graphic imaging, to name a few. It has also created a world in which there is greater need for humans to communicate across the barriers of language, geography and time. As work in telecommunications research attempts to expand communication media beyond the telephone, it faces a vast dilemma: what is the best configuration of available devices for overcoming these barriers and facilitating human-to-human communication?

One logical place to begin is with the effort to understand how humans behave in the course of communication with one another in the absence of technological aid. It is then instructive to compare this behavior with the performance of humans communicating in a machine-mediated environment. Such an environment has been created at ATR by Loken-Kim, Kitagawa, and others: EMMI, the ATR Environment for Multi-Modal Interaction (Loken-Kim *et al.*, 1993a). The work reported here is part of an investigation of linguistic and communicative behavior of humans in the EMMI environment. It describes the results of an examination of client/agent interaction when engaged in a directions task via both telephone and EMMI .

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 Giles *et al.*, 1987 for a survey). For example, Oviatt (1994) 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 results are not surprising in that the form-based condition was extremely restrictive, severely limiting the conversants' choices, and thus their opportunities for disfluency.

The work reported on here was 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.

There was a more general aim to this work as well. What are the effects of the mode of interaction on conversants' goal-directed behavior? In particular, are conversants more efficient in achieving informational goals in one mode than in another? Further, does their linguistic behavior reflect differences in their attitudes about how accessible or "comfortable" each mode is?

Methods

Eight subjects, all native speakers of American English with no significant problems involving sight or hearing, 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 EMMI. They were encouraged to be as natural as possible and to extend the conversation as long as necessary for them to feel comfortable with the information they received. (The full instructions appear in Loken-Kim 1993b.) Four subjects participated in the telephone situation first; four used EMMI first.

The agent in all trials was a trained, native speaker of American English. The agent, and thus, to some extent, the agent's speech, was kept constant so that client speech patterns would not be affected by interaction with different speech patterns of different agents. 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). The headset allowed the conversation to be taped in the same way that the audio taping was accomplished in the multi-media (MM) environment. The client and agent talked and listened to one another through the telephone.

In the MM environment, subjects spoke into the same headset-mounted microphone, 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). The full description of EMMI can be found in (Loken-Kim *et al.* 1993a). 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. Subjects were videotaped from the front to record facial and head movements and from the side to record manual movements. (In addition, the front view video provided the image that appeared on the agent's monitor.) Agents were also videotaped from the front (this image appeared on the clients' monitor). 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; blatant deviations from standard pronunciation[1]; the pronunciation of "the" as /thi/ and of "a" as /e/; and simultaneous speech. The transcriptions were checked twice, all by independent transcribers. The full set of transcriptions is given in (Loken-Kim *et al.*, 1993b).

Measures

Disfluency. The standard measures for disfluency include the number of false starts made and and the number of filled pauses used per 100 words[2]. While the latter were fairly easy to identify, the former required a judgment call in some cases, taking into consideration intonation and context. Where there was a question, the three transcribers discussed the case until a consensus was reached. In order to assess disfluency, the measures for false starts and for filled pauses were added, and the disfluency rate for each participant for each conversation was calculated per 100 words.

Efficiency. Measures relevant to the definition of "efficiency" were more problematic. Clearly the notion of an "efficient" conversation implies that the conversation is a means to achieve one or more goals; one measure of "efficiency," then, could be the amount of time

[1]We noted sounds that were perhaps truncated from intended utterances such /l/ or /th?/. We transcribed as words utterances such as "wanna" and "gonna." Certainly we were influenced by the fact that there is an available and standard orthography for these examples; the equally common, but usually unnoted, /we:r/ for "where are" was noted similarly to the truncated utterances above, but could as easily have been rendered as a word if there had been a standard orthographic form for this expression. To some degree, then, the distinction between "word" and "deviant pronunciation" is blurred and arbitrary.

[2]A measure that more accurately reflects "conversational fluency" rather than production fluency would involve an examination of the large amount of simultaneous speech found in this corpus. An (intuitively) fluent conversation involves a fair amount of simultaneous speech, possible because the two speakers can predict each other's utterances accurately enough to pre-empt their endings (Fais, in press). Occasionally, simultaneous speech results in the repetition of the piece contributed by the speaker whose turn is next, or may also result in complete back pedalling to ensure that a contribution has been understood. These cases are examples of conversational disfluency. This corpus has not yet been analyzed for this type of fluency.

taken to achieve some goal. We labelled the transcribed conversations for discourse goals pursued and found that these goals fell into two categories. In the first are standard kinds of discourse goals, in contexts that would be expected in this type of task: primarily clients seeking information related to the task ("how much money do I need for the bus?") and the agent seeking information relevant to performing her function ("where are you calling from?"). These were called "solicited goals." The second type of goal we called "offered goals;" these were cases in which the information was not requested but had been offered by either agent or client (usually the former, e.g., "The bus ride will cost you five hundred yen [answer to request for information] and will take a half an hour [offered goal]"). We then were able to assess the average time required per goal, as well as average words and average turns per goal for each conversation.

Comfort. No matter how fluent or efficient a mode or configuration of modes is, a communication environment is not useful unless it is also accessible to naive users. An objective measure for "comfort" is even more difficult to contrive than that for "efficiency;" however, we did administer a questionaire designed to elicit clients' reactions to the two communication environments, whose results shed some light on this issue. The responses to that questionaire are summarized in the Discussion below. In addition, with respect to this point, we also investigated the number of words used per speaker, and number of goals achieved for each conversation. We reasoned by analogy to human-to-human conversation. If speakers are comfortable talking together, they tend to talk longer and pursue a greater number of topics than if they are uncomfortable. Similarly, we propose, if speakers are more comfortable in a particular communication setting, they will use more words and will attempt more goals per conversation than in an uncomfortable setting.

Results

Disfluency. As seen in Table 1, there was no significant effect of mode alone on disfluency. However, clients' disfluency rates were significantly affected by the interaction of mode and order. Clients who participated in the telephone condition first, followed by the MM condition, had fewer disfluencies in *both* those environments than did clients beginning with MM followed by telephone in *either* environment (Figure 1). In other words, it seems that the disfluency rates typical of telephone speech were lower than those typical of speech in the MM environment. Furthermore, the rates typical to the first mode used by the client, regardless of which mode that was, tended to persist into the second mode as well. As a result, the disfluency rate for telephone mode was elevated following MM mode, and that for MM mode was lowered following telephone mode.

ANOVA Table for disfluency
Split By: ag/cl
Cell: C

	DF	Sum of Squares	Mean Square	F-Value	P-Value
mode	1	2.083E-4	2.083E-4	.112	.7435
order	1	.002	.002	.882	.3661
mode * order	1	.010	.010	5.306	.0400
Residual	12	.022	.002		

Table 1. ANOVA table for client disfluency rates with respect to mode, order and mode interacting with order.

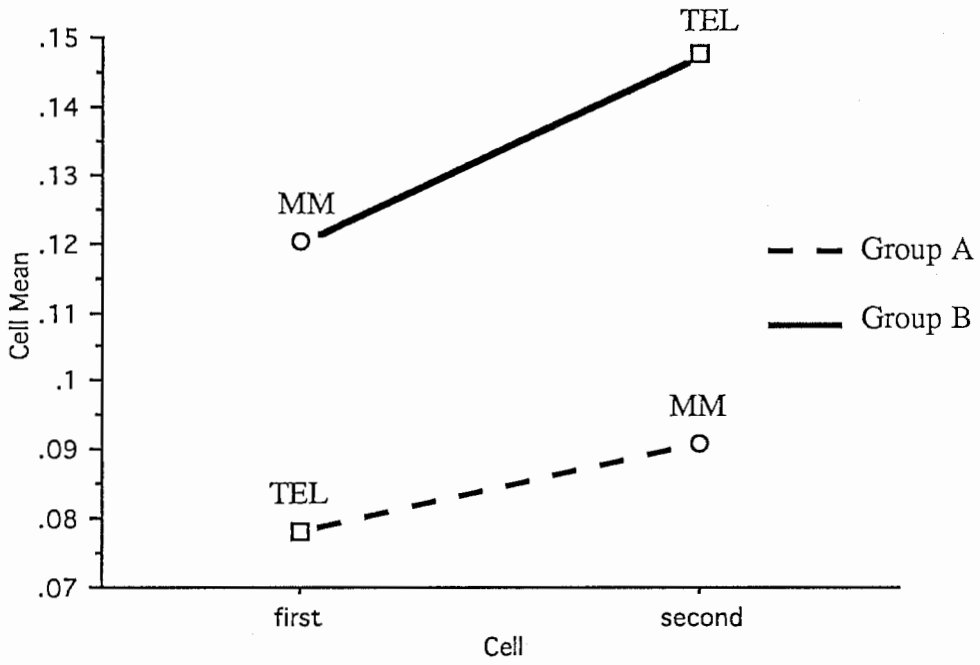
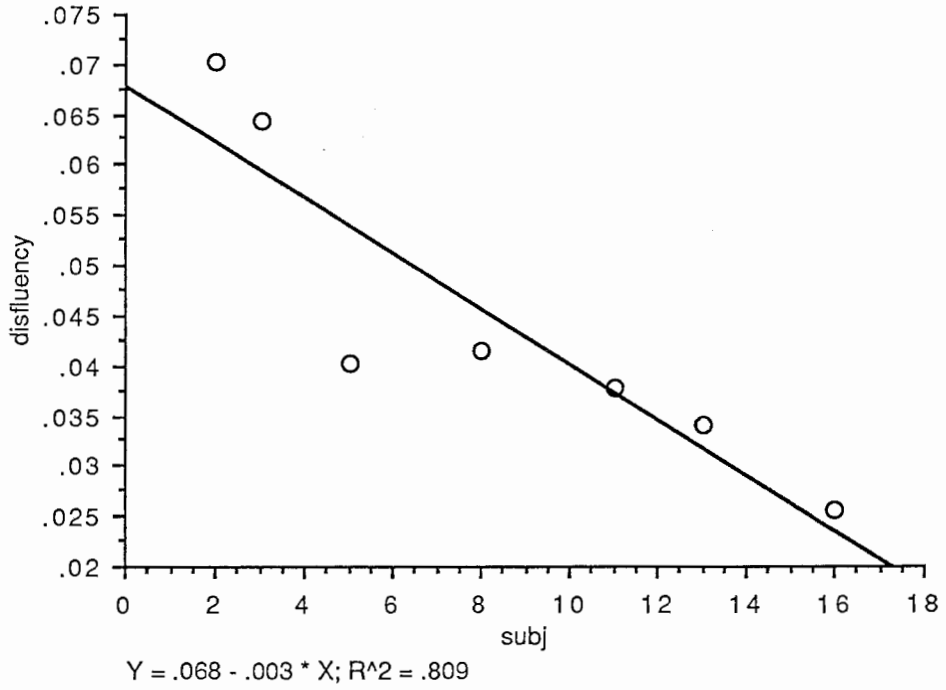


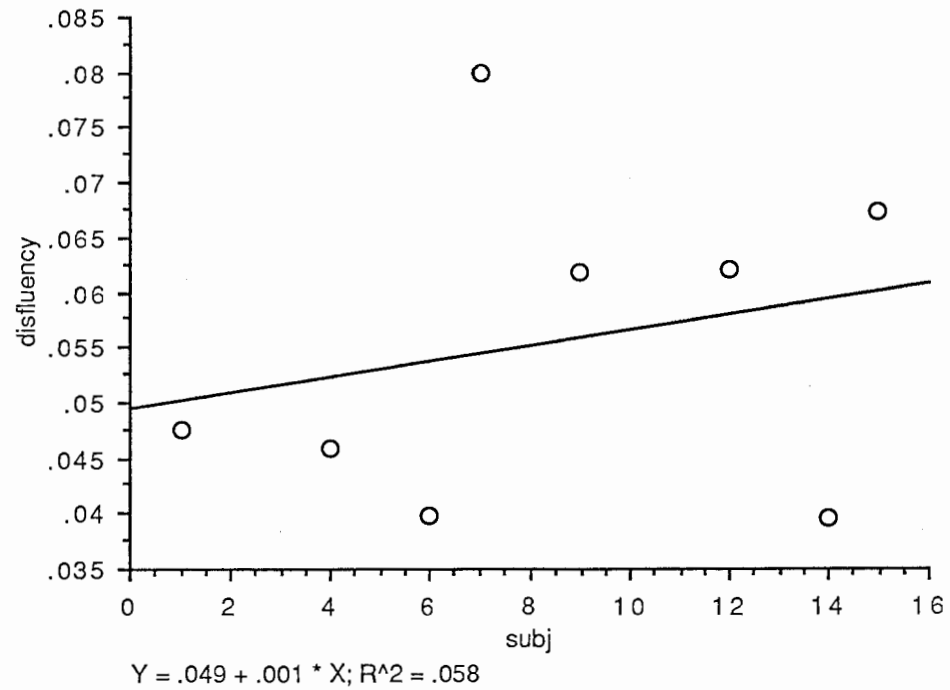
Figure 1. Interaction plot for client disfluency with respect to mode interacting with order.[3]

The agent's speech cannot be indicative of trends in disfluency as influenced by mode or order since there was only one agent in the experiment. However, it is interesting to observe the changes in the disfluency rates of the agent over time. When plotted against the serial order of trials, the agent's speech in the MM environment shows a marked tendency

[3]The disfluency rate of around 8% for the telephone first condition, is similar to the rate cited in Oviatt (1994) for a two-person telephone call (8.83%). Note that the disfluency for the MM first condition is

to become less disfluent (Figure 2), while her telephone speech shows no such trend (Figure 3).





almost double that for the telephone first condition.

Efficiency. The initially proposed measures for efficiency were time per goal, turns per goal and words per goal. The first two measures may not, in fact, necessarily be indicative of efficiency: speakers may speak more quickly or more slowly regardless of efficiency, and may take a smaller number of longer or a greater number of shorter turns, again without affecting efficiency. The number of words used to achieve a goal is perhaps the most reasonable measure of the effective achievement of goals. In fact, clients and agent did not use significantly less time or significantly fewer turns to achieve their goals in any particular experimental condition. However, there *was* a significant difference in the number of words used by the client per goal in some conditions (Table 2). This difference was affected by order only; clients became more efficient, as might be expected, in the second condition, regardless of which mode was employed first and which second (Figure 4).

ANOVA Table for words/goal
Split By: ag/cl
Cell: C

	DF	Sum of Squares	Mean Square	F-Value	P-Value
mode	1	5.473	5.473	1.015	.3336
order	1	26.061	26.061	4.832	.0483
mode * order	1	6.465	6.465	1.199	.2951
Residual	12	64.722	5.394		

Table 2. ANOVA table for client words per goal with respect to mode, order and mode interacting with order.

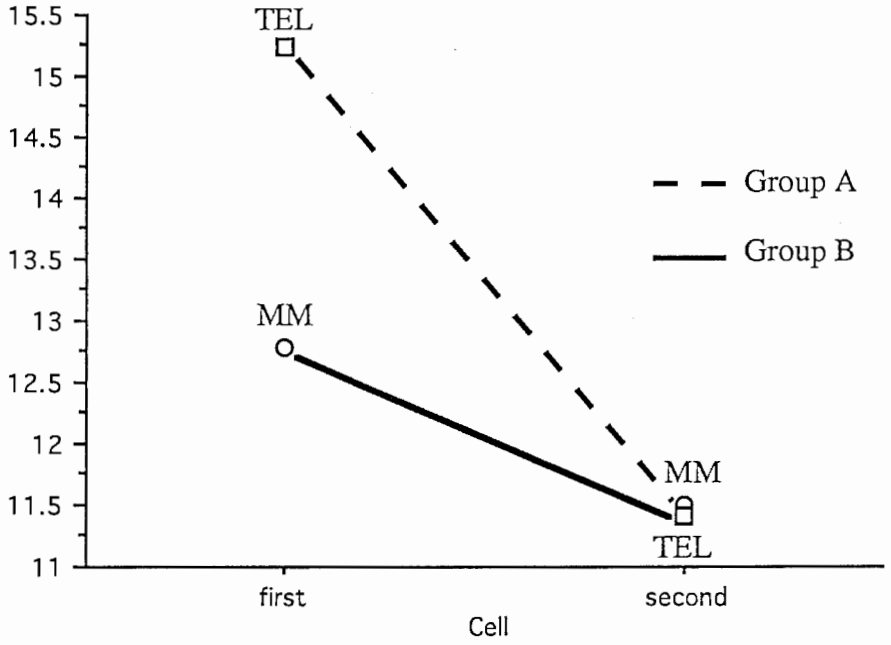


Figure 4. Interaction plot for client words per goal with respect to mode interacting with order.

As remarked above, the agent's speech is not a reliable indication of trends since it was produced by only one person. However, the agent's numbers of words per goal showed an interesting pattern, which we mention here for the sake of the discussion in the next section (Figure 5). Like the clients' disfluency rates, the words per goal of the agent differed depending upon both mode and order of condition. Again, like client disfluency rates, the agent's words per goal were lowest in the telephone first condition, and increased in the next, MM, condition. Similarly, when the MM condition was first, her words per goal were the highest, and she became more efficient in the following telephone condition, though not as efficient as in the telephone first condition. When the telephone condition was first, the greater efficiency of that condition persisted into the MM condition, and the MM condition following the telephone condition was more efficient than the MM first condition. On the other hand, he efficiency of the telephone condition was lowered when that condition followed the MM condition.

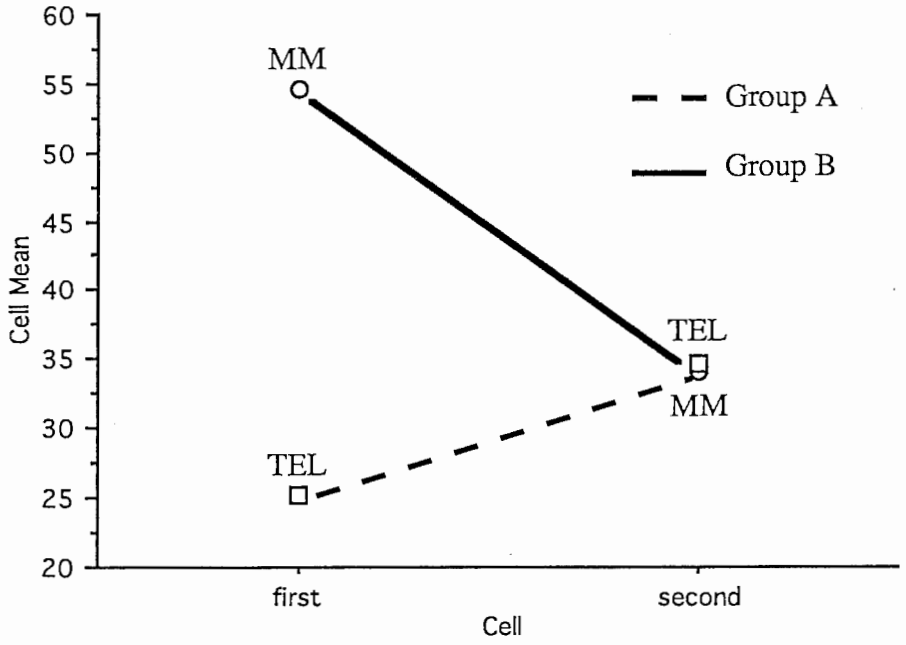


Figure 5. Interaction plot for agent words per goal with respect to mode interacting with order.

Comfort. As discussed above, the measures proposed for comfort were number of words used by the client and number of goals accomplished. The distinction between sought goals and offered information turned out to be a significant one; offered information

showed no mode or order effects, while sought goals were significantly affected by the interaction between mode and order. The results for number of words used by the client *and* for number of sought goals in the conversation followed the same pattern as client disfluency rates. That is, clients who used the telephone first and MM second used more words and accomplished more goals in *both* the telephone and MM environments than did clients beginning with MM and then using telephone in *either* environment. Significance results for words are shown in Table 3; quantitative results for words are illustrated in Figure 6. Significance results for goals are shown in Table 4; quantitative results for goals are illustrated in Figure 7.

ANOVA Table for words
Split By: ag/cl
Cell: C

	DF	Sum of Squares	Mean Square	F-Value	P-Value
mode	1	1122.250	1122.250	.302	.5928
order	1	484.000	484.000	.130	.7245
mode * order	1	36864.000	36864.000	9.915	.0084
Residual	12	44615.500	3717.958		

Table 3. ANOVA table for client words with respect to mode, order and mode interacting with order.

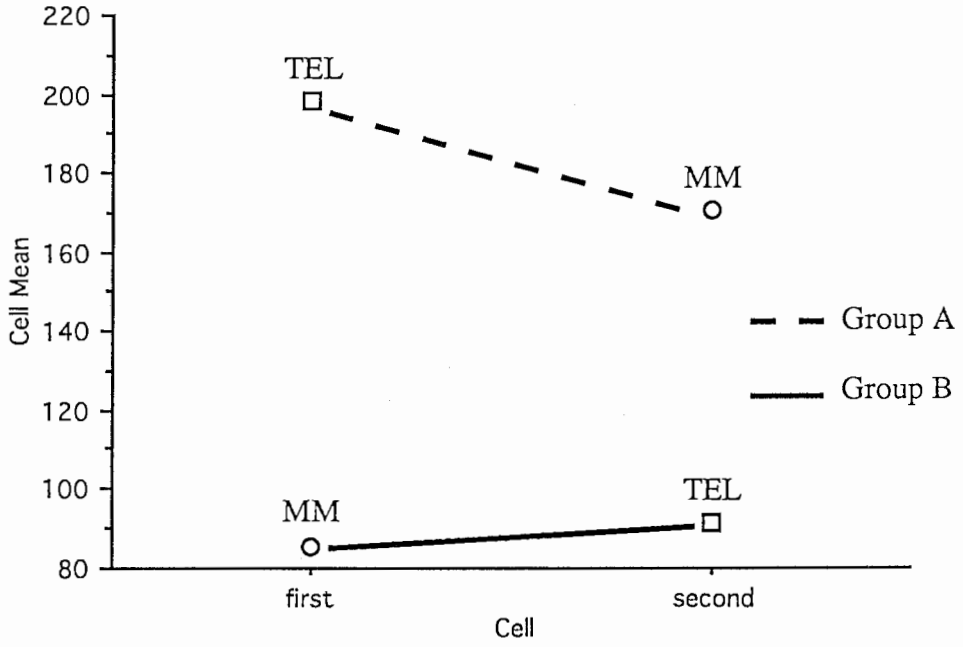


Figure 6. Interaction plot for client words with respect to mode interacting with order.

ANOVA Table for sought goals

	DF	Sum of Squares	Mean Square	F-Value	P-Value
mode	1	2.250	2.250	.248	.6277
order	1	6.250	6.250	.688	.4230
mode * order	1	90.250	90.250	9.936	.0083
Residual	12	109.000	9.083		

Table 4. ANOVA table for sought goals with respect to mode, order, and mode interacting with order.

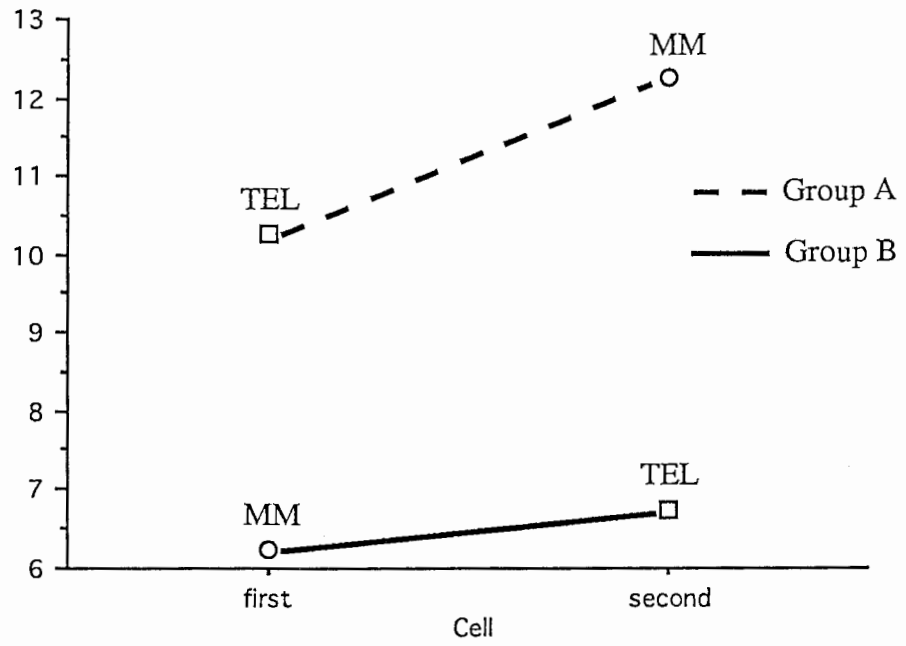


Figure 7. Interaction plot for sought goals with respect to mode interacting with order.

The agent's speech showed an interesting trend over the course of the experiment. In the MM condition only, as the experiment proceeded, she used fewer words (Figure 8). In the telephone condition, there was no patterned difference over time in her use of words (Figure 9).

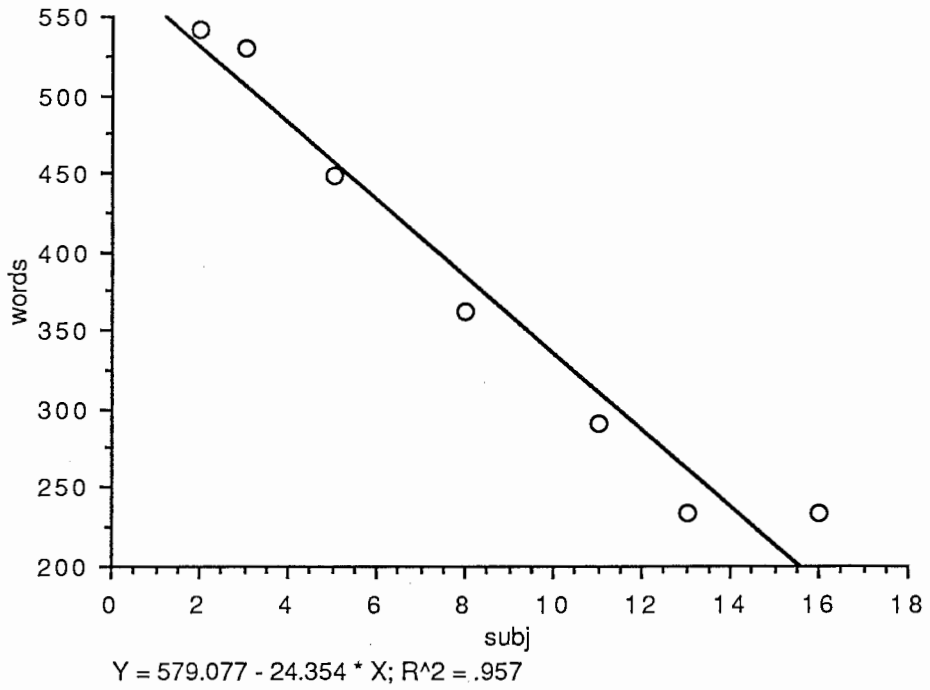


Figure 8. Regression plot for agent words in the MM condition over serial order of trials.

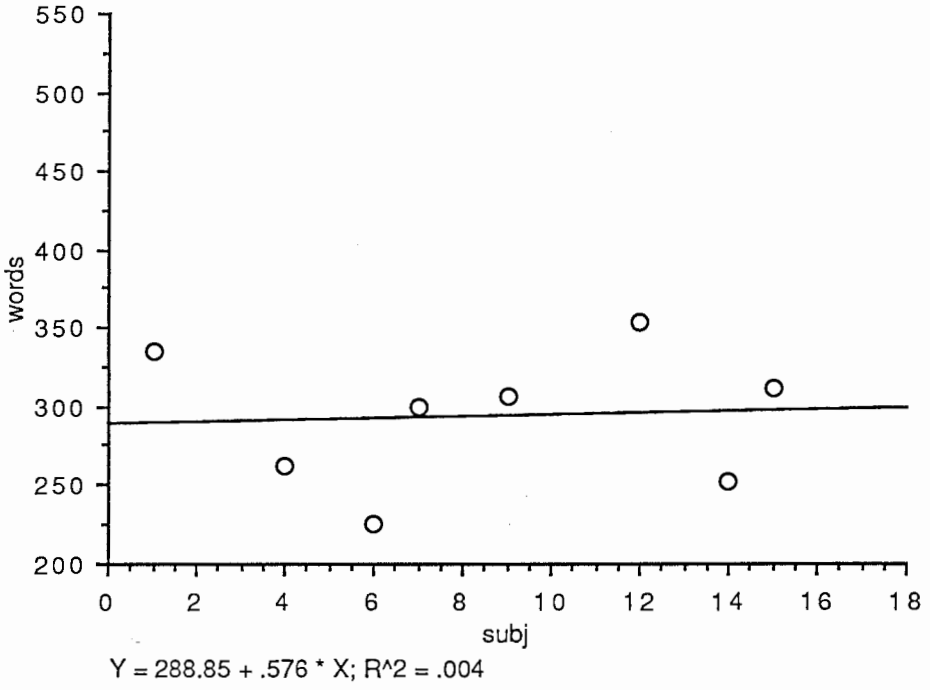


Figure 9. Regression plot for agent words in the telephone condition over serial order of trials.

Discussion

Clearly, in terms of client disfluency rates, and comfort (number of words and number of goals), the key issue seems to be in which order the client encounters the two conditions. We propose to explain these results on the basis of familiarity: because the clients are more familiar with the telephone, they show a greater comfort level and less disfluency in the telephone-initial conditions. On the other hand, when they experience the MM condition first, that being a completely unfamiliar environment, they show the highest levels of disfluency and the lowest levels of comfort. These behaviors then persist into the following condition: if the following condition is the MM condition, subjects retain some of the comfort and fluency of the telephone condition even in that unfamiliar environment. If the second condition is the telephone, subjects carry over a higher disfluency and less comfort to this familiar environment from the unfamiliar initial condition.

For the agent, on the other hand, the MM environment was much more familiar; she had ample opportunity to practice with the system before the experiment began and, of course, her experience with it accumulated during the course of the experiment. Thus, the agent showed none of the sort of carry-over effects between familiar and unfamiliar environments that clients showed in her disfluency rates or number of words used. The pattern of the agent's behavior was quite different. She showed a decrease in disfluency rates in the MM condition *only*, across the course of the experiment. We interpret this as meaning that, even though the MM environment was familiar to the agent, she, too, was showing the effects of practice with the environment in that she was becoming more fluent in it.

The differing patterns for client and agent were strikingly corroborated by the goals results. Recall that we distinguished two types of goals: sought goals and offered information. We did this because it was clear from the data that informational goals were being achieved even when they were not initiated by the client. Thus, even though the accomplishment of informational goals is a joint achievement of agent and client, we are still able to differentiate agent and client behavior with respect to goals: "sought goals" are, in some sense, client goals, while "offered information" constitutes agent goals. In fact, these two types of goals showed exactly the same patterns as client and agent disfluency rates and number of words: client goals (sought goals) showed carry-over effects due to mode and order while agent goals (offered information) did not.

These results indicate that the simple fact of a multi-media environment does not influence clients' behavior sufficiently to render their speech more fluent and more machine-processable. On the contrary, the subjects' lack of familiarity with the MM environment rendered their speech less fluent, and less processable. Thus, the lower disfluency rates

shown by Oviatt (1994) in form-based exchanges must have been due, not the the environment in which the exchanges took place, but to the organization of the task itself.

The agent also tended to use fewer words in the MM environment over the course of the experiment. While we have equated fewer words with less comfort in the case of the client, we would interpret the case of the agent differently. The agent always used far more words than the clients did, due to the nature of her role as information giver, and possibly also because she was more comfortable than the client in the experimental environment (having had more experience with it). In addition, she used many more words in the MM condition than in the telephone condition. In fact, then, the use of fewer words per MM conversation by the agent may indicate that she was becoming more efficient in conveying information in the MM environment. That this is in fact attributable to increasing familiarity with the media can be inferred from the fact that there was no similar trend for the agent over the telephone conditions.

Client efficiency, on the other hand, seemed to be affected only by practice and not by mode. Clients seemed to approach the tasks with about the same efficiency regardless of whether they were conversing on the telephone or in the MM environment, and they simply got better the second time around. One mode of interaction did not evoke more or less efficient responses than the other mode.

The case of agent efficiency is different. The agent showed the same carry-over effects of familiarity in her efficiency (words per goal) that the client showed for disfluency and comfort. We interpret the agent's case in this way. A comparison of the telephone-first and MM-first efficiency rates showed that the MM environment is less efficient. This is not surprising; in the MM condition, the agent uses a greater number of words: to inform the client of the actions she is taking to show a map on the screen, to "fill time" while the map appears, and to check that the client can see and understand the visual image. For the agent, the speech patterns set in the first condition carried over to the second condition, rendering the telephone second condition more verbose and the MM second condition less verbose than their respective first conditions. The carryover in the case of the agent, then, is a result of the qualitative differences in the linguistic tasks the agent must perform in each of the two media, rather than familiarity level.

What does this tell us about the design of multi-media interfaces? The multi-media environment seems to have greater potential for more efficient conveyance of information, given the fact that it can draw on the visual as well as the auditory modes in a number of different ways. However, the results from just this pilot study suggest that users are not familiar enough with such configurations to use them to their full potential. On the other

hand, users have had extensive experience with the technically less efficient instrument, the telephone, and results suggest that that familiarity and practice with that medium completely offset the greater design advantages of the MM environment.

Clearly this limits the contributions that a MM environment can make in a machine translation situation. Rather than reduce disfluency, multi-media options in fact increased disfluency, rendering the speech of the clients in this environment *less* susceptible to automatic processing.

On the other hand, if MM environments such as this one are to be incorporated into the public service sector as in the scenario in this experiment, then the impressions of the clients are also of importance. In this regard, it is instructive to note the results of the post-experiment interview conducted with each of the subjects.

Immediately after the conclusion of the experiment, subjects were asked to fill out a post-experiment interview questionaire designed to solicit such impressions as ease of use, comfort or usefulness of each of the modes of interaction. The subjects rated these parameters on a scale provided in the questionaire by marking an X somewhere along the scale. The responses to this task as given by the subjects on these questionaires are represented collectively below.

First, each subject rated how enjoyable the experiment had been:

Telephone:

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

Multi-media setting:

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

Clearly the greater familiarity with the telephone was a detriment in terms of how much fun it was to use. While this is perhaps not a crucial issue, it is at least instructive to note that the lack of familiarity with the MM environment did not lead to disinterest.

Of more pertinent interest is the rating the subjects made as to how easy it was to use the different media:

Telephone:

__XXXX____XXXXXX_____

simple some effort had to work at it difficult

Multi-media set-up:

XXXXXX_____XXXX_____

simple some effort had to work at it difficult

Despite the conjecture that lack of familiarity with MM might account for greater client disfluency rates and lower numbers of words used and goals achieved, the subjects themselves did not perceive a greater amount of difficulty in using the MM set-up. This might have been affected by the greater enjoyment in using the MM environment which they expressed above.

Up to this point, we have been discussing the MM environment as a single entity. In fact, it involved several different media in combination: the map, keyboard-mediated typewritten expression, and the video image, as well as the audio channel. Clients perceived the usefulness of these different media in very different ways:

Telephone:

__XXX_____XXXXXXX_____

very useful served some purpose an inconvenience worthless

Map:

_XXXXXXXX____XX_____

very useful served some purpose an inconvenience worthless

Keyboard:

__X_____XXXXXXXXX____

very useful served some purpose an inconvenience worthless

Clearly, subjects recognize the advantage of the visual information contained in the map, even though their linguistic performance was influenced by other factors. On the other hand, they also recognized that the telephone allowed a somewhat more restricted communication of information. In addition, since very few of them used the keyboard at all, that aspect of the MM environment was rated very low.[4]

Of course, the linguistic results achieved are also influenced by the nature of the task involved. It was hoped that speakers would naturally edit and "clean up" their speech in the MM environment and that, for that reason, no further constraints would have to be placed on their conversation in order to render it maximally suitable for automatic language processing. For that reason, the task in this experiment was left quite open-ended. Clearly that hope was far too optimistic. The effects of mode alone on fluency, in fact, were the reverse. Unlike Oviatt's form-based condition, in which the *task* so restricted the subjects that their speech became more fluent, *mode* was not sufficient to produce the same results. It still might be possible, however, to redesign the task involved so that it better exploits the advantages of the MM environment while at the same time not providing as restrictive an activity as that found in Oviatt (1994). This is an issue for future work in the ATR EMMI.

Bibliography

Fais, Laurel, in press. "Conversation as Collaboration: Some Syntactic Evidence." *Speech Communication.*

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

Loken-Kim, Kyung-ho, Fumihiro Yato, Kazuhiko Kurihara, Laurel Fais, and Ryo Furukawa, 1993a. EMMI - ATR environment for multi-modal interactions. ATR Technical Report TR-IT-0018. Kyoto, Japan: ATR Interpreting Telecommunications Research Laboratories.

[4]The lack of use of the keyboard was a result of two factors. First, at no time in the task was the use of the keyboard crucial to the achievement of goals. Second, again, the lack of familiarity with the MM options tended to make subjects deal only with what was essential and useful.
A summary of the suggestions and comments of the subjects is given in (Loken-Kim 1994), and available from the author upon request.

Loken-Kim, Kyung-ho, Fumihiro Yato, Laurel Fais, Kazuhiko Kurihara, Ryo Furukawa, Yoshihiro Kitagawa, 1993b. 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.

Loken-Kim, Kyung-ho, Fumihiro Yato, Laurel Fais, Tsuyoshi Morimoto, 1994. Linguistic and paralinguistic differences of telephone-only and multi-modal dialogues. ICSLP, Yokohama, September, 1994.

Oviatt, S.L., in press. Predicting spoken disfluencies during human-computer interaction. Proc. CHI '94, Boston.