6th International Conference on Spoken
Language Processing (ICSLP 2000)
Beijing, China
October 16-20, 2000

PREFERRED MODALITIES IN DIALOGUE SYSTEMS

Vildan Bilici, Emiel Krahmer, Saskia te Riele and Raymond Veldhuis

IPO, Center for User-System Interaction,
Technische Universiteit Eindhoven, The Netherlands
{V.Bilici/E.J.krahmer/S.M.M.t.Riele/R.N.J.Veldhuis}@tue.nl

ABSTRACT

This research describes which modalities are preferred in particular contexts when interacting with a multi-modal dialogue system. The trade-off between three factors is investigated: (*i*) speech recognition performance, (*ii*) efficiency of input modality and (*iii*) the system's output modality. Four versions were developed of a multimodal examinator to be used in elementary school. The versions differed in recognition performance ('perfect' vs. realistic) and output modality (speech or text). In all systems, subjects could provide input via speaking or typing. Answer length in characters was used as a measure of efficiency. Results show that both speech recognition performance and efficiency have a strong impact on preferred modalities. No effect was found of the system's output modality.

1. INTRODUCTION

"Speech is the bicycle of user-interface design," according to Shneiderman (1998:328), "it is great fun to use (…), but it can carry only a light load. Sober advocates know that it will be tough to replace the automobile: graphic user-interfaces." The 'bicycle' status of speech is probably due to two factors. First, correcting speech recognition errors in spoken mode is difficult and reduces user satisfaction (see e.g., Weegels, 1999, Krahmer et al. 2000, Swerts et al. 2000). A second limitation is that for some tasks (such as presenting lists) speech is a sub-optimal modality. Nevertheless, Shneiderman's statement is misleading in that it suggests that speech and graphical interfaces are mutually exclusive. This is certainly not the case, as shown by the recent emergence of sophisticated multimodal interfaces which combine the advantages of speech, graphics and other modalities (see e.g., Oviatt and Cohen 2000).

For the development of efficient multi-modal interfaces, it is important to know why people opt for certain (combinations of) modalities in a given context. The literature contains various different claims which are relevant for this question. For instance, Oviatt and coworkers (e.g., Oviatt and Olsen 1994; Oviatt et al. 1998) found that when users experience difficulty giving input via one modality, they tend to switch to a less error-prone input-channel. According to Oviatt and Olsen (1994) this is an instance of what they call "contrastive functionality." Another instance which Oviatt and Olsen distinguish is the difference between digits and text. They found that digits are less likely to be spoken than ordinary text, irrespective of the length of the numerical input. This latter finding runs counter to the claim from Baber (1991) that length of the input does have an influence. In particular, he found that *short* input (limited number of characters) is more efficiently entered via keyboard, and arguably this may influence modality selection. This might be related to the general tendency in humans (observed by e.g., Zipf 1949) to opt for minimal effort.

The experiments of Oviattt et al. were done with a simulated "Service Transaction System", which can assist users with such tasks as renting a car or personal banking. Subjects had the possibility of to enter information by voice or by writing (on an LCD tablet). It is worth noting that subjects did not really engage in a *dialogue* with the service system, in the sense that subjects were not explicitly prompted for information by the system. It seems a reasonable hypothesis that the modality in which subjects are prompted has an influence on their own choice of modality. This hypothesis is based on the work of Reeves & Nass (1996), whose main tenet is that humans treat computers as social actors. For example, it is well-known that most humans are polite to each other most of the time. Reeves & Nass show that humans also have a tendency to be polite to computers. One politeness rule which they mention is the *Rule of Matched Modality*. This rule is based on the (informal) observation that it is polite to respond to someone using the same medium (e.g., a letter should get a letter in response). Reeves & Nass note that in many current computer products there is an asymmetry in that users cannot always respond using the same modality. They write: "If an interface accepts only text input, perhaps it should produce only text output. If the user can respond with voice, then a voice-based interface might work better." This implies that when users can *choose* between text or speech, they are more likely to give their input in the same modality as the system gave its output.

In this paper the trade-off between the aforementioned factors contributing to modality selection (error avoidance, efficiency and matched modality) is studied in the context of multi-modal dialogue systems, using a specially designed automatic examinator.

2. METHOD

As a research platform, a multi-modal simulation of an automatic examinator (Dutch) was developed, to be used in elementary school. The main advantage of such an examinator is that it corresponds

Table 1: Percentages of spoken and typed answers.

	system variant			
input modality	TEXT^{+}	SPEECH^{+}	TEXT^{-}	SPEECH^{-}
speech	88%	85%	65%	62%
text	12%	15%	35%	38%
total	100%	100%	100%	100%

Table 2: Numbers of modality switches from speech to text, and their sources.

	system variant			
speech → text	TEXT^{+}	SPEECH^{+}	TEXT^{-}	SPEECH^{-}
ASR error	-	-	44	60
char to digit	9	4	1	4
other	5	1	4	18
total	14	5	49	82

Table 3: Numbers of modality switches from text to speech, and their sources.

	system variant			
text → speech	TEXT^{+}	SPEECH^{+}	TEXT^{-}	SPEECH^{-}
post ASR error	-	-	37	53
digit to char	7	4	3	6
other	12	0	9	20
total	19	4	49	79

closely to a common dialogue situation (of the question-answer variety) in which both spoken and written input are highly natural and simple. At all times, subjects could provide their answers by voice or by typing. In the former case they could speak into the microphone standing next to the computer terminal and press the mouse-button to indicate they finished talking. In the latter case they could enter their answer in a text-field in the middle of the screen and had to press enter or the mouse button afterwards.

Four versions of the examinator were constructed. The systems differed in (a) their recognition performance, and (b) their output modality. The recognizer either performed perfectly (no errors) or realistically (20% word error rate). The perfect recognizer was simulated by simply accepting all the answers as correct. In the realistic recognizer, items to be misrecognized were selected randomly. Half of these were immediately rejected. In the remaining cases the system indicated that it could not understand the subjects' answer and offered a second or even a third chance, which only lead to success in a limited number of cases. The systems either used speech or text as their output modality. In the systems which used visual text output, the questions were posed and the answers evaluated on the screen. In the other systems, the interaction was done using a pre-recorded male voice. We opted for a male voice partially because praise from males is generally taken more seriously than praise from females (Eagly and Wood 1982). Pre-recorded speech was used in an attempt to maximize the perceived communicative abilities of the automatic examinator. This was done to make the perfect performance of the recognizer more plausible. In the remainder of this paper, the four versions will be designated as TEXT^{+}(text output / perfect speech recognition), SPEECH^{+}(speech output / perfect speech recognition), TEXT^{-}(text output / realistic speech recognition) and SPEECH^{-}(speech output / realistic speech recognition).

During the experiment, subjects were observed through a one-way screen. Before the actual experiment subjects had to utter a test sentence to simulate microphone calibration. This was done to lead subjects to believe real recognition was performed. The experiment started with a short introduction in which subjects had to answer 5 questions with each input modality. Half of the subjects first had to answer vocally, the other half typing. In the systems with a realistic recognizer one answer was misrecognized during training. After the training session, the examinator took 7 exams of 10 questions each (about arithmetic, history, topography etc.), all priming subjects to give a one-word answer (for example: "In which country is the Chinese Wall located?"). It was decided to use questions at the level of elementary school to be reasonably sure that the vast majority of subjects would be able to answer nearly all questions correctly, which was indeed the case (though one subject thought the Chinese Wall was to be found in Berlin). This makes automatic acceptation of answers possible without actually checking the input. For each version, 22 questions could be answered using at most three (numerical) keystrokes (for instance, "How much is 10 times 11?"). Subjects were told that a book token of 25 euro would be awarded to the best performing subject. During the experiment, subjects received continuous feedback about the number of correct answers and whether their performance so far was above or below average. This average was set at 9.1 so that due to one misrecognized yet correct answer per exam, the subject immediately scored below average. This was done to induce a comparable decrease in user satisfaction caused by recognition errors as found in operational dialogue systems. After the experiment, subjects filled in a questionnaire, in which they were asked which input modality they preferred and why. Finally, they were informed about the absence of actual answer checking during the experiment, and the falsely accepted answers (if any) were discussed. Given the predestined outcome of the exams, the book token was raffled.

In the experiment, 56 (+ 2) subjects participated; 14 for each version of the multi-modal examinator. Two subjects found out during the experiment that they were working with a simulated system that performed no real automatic speech recognition; their results were not taken into account. Subjects were mostly students (a sizeable amount studying to become elementary school teachers) who volunteered. All were native speakers of Dutch. Roughly half of the subjects had some experience with speech recognition, and all could type with at least two fingers.

3. RESULTS

If the rule of matched modality applies, we would expect subjects to type more often in the TEXT versions and to use speech more often in the SPEECH versions. In addition, following Oviatt and Olsen (1994), it seems plausible that in the $-$ versions (realistic recogni-

Table 4: Subjective modality preference

	system variant			
pref. modality	TEXT^{+}	SPEECH^{+}	TEXT^{-}	SPEECH^{-}
speech	13	12	8	8
text	1	2	6	6
total	14	14	14	14

Table 5: Initial (I) and final (F) modality.

	system variant							
input	TEXT^{+}		SPEECH^{+}		TEXT^{-}		SPEECH^{-}	
modality	I	F	I	F	I	F	I	F
speech	9	14	13	12	10	10	10	8
text	5	0	1	2	4	4	4	6
total	14	14	14	14	14	14	14	14

Table 6: Numbers of subjects using either speech or text or both as input during their interaction with the examinator.

	system variant			
input modality	TEXT^{+}	SPEECH^{+}	TEXT^{-}	SPEECH^{-}
speech	6	10	1	0
text	0	1	2	2
speech+text	8	3	11	12
total	14	14	14	14

Table 7: Percentages of typed digits and non-numerical answers.

	system variant			
	TEXT^{+}	SPEECH^{+}	TEXT^{-}	SPEECH^{-}
non-digits	6%	14%	33%	32%
digits	26%	17%	37%	46%

tion) subjects type more often than in the $+$ version ('perfect' recognition). Table 1 shows the percentages of spoken and typed answers. When recognition is perfect, 88% (TEXT^{+}) and 85% (SPEECH^{+}) of the answers are spoken. When recognition is realistic these percentages drop to 65% and 62% respectively. This effect of recognition performance is significant in an ANOVA ($F_{(1,52)} = 14.858$, $p < .001$). This supports the claim that people look for alternative modalities in case of errors. The difference between the TEXT and SPEECH versions, however, is not significant ($F_{(1,52)} = .01$, $p = .917$). This provides no evidence for the rule of matched modality. It is interesting to note that subjects overall use text substantially more often than reported in Oviatt and Olsen (1994). It might be that typing is more efficient than writing on an LCD tablet.

Table 2 lists the number of times subjects switch from speaking to typing, and also gives the respective causes. Automatic speech recognition (ASR) errors are indeed the primary source of modality switches. Additionally, in the majority of the cases, subjects switch back immediately after having done with the problematic item (see table 3). Apparently speech is more efficient and easy to use, as subjects in all four conditions explicitly wrote in the questionnaire. The efficiency of speech in the current experimental paradigm can probably be attributed to the finding that the translation of thought to speech is much faster than the translation of thought to typing (Chalfonte *et al*. 1991). A further indication of the impact of efficiency on modality selection may be derived by comparing digits and text. Overall subjects are more likely to answer questions about arithmetics via typing (which always required at most three keystrokes) than other questions (which generally require substantially more keystrokes): 32% of the numerical questions is answered by typing as opposed to 22% of the non-numerical questions. This difference is significant ($t_{(55)} = 2.451$ (paired), $p < .05$), suggesting that typing becomes more efficient as the input gets shorter. However, one should also keep in mind that subjects still use speech for 68% of the numerical data. Given the perceived overall efficiency of speech, one would expect that subjects also prefer speech. However, subjects only have a significant preference for speech when they work with either the TEXT^{+}($\chi^2_{(1)} = 10.28$, $p < .01$) or the SPEECH^{+}variant ($\chi^2_{(1)} = 7.14$, $p < .01$); see table 4.

So far, the results provide no evidence for the rule of matched modality. Even though subjects predominantly speak in the SPEECH conditions, this is clearly not an effect of matching modalities, since speech dominates in all *four* conditions. This implies that if an effect of the rule of matched modalities is to be found at all, it should be in the TEXT conditions. For instance, in the TEXT condition subjects might *start* typing, but switch to speech during the course of the experiment. Table 5 shows that this is not the case. It could also be that the number of subjects which type at least *some* of the times is significantly higher in the TEXT than in the SPEECH conditions. Again this is not found in the data, as can be seen in table 6. Although it is true that more subjects in the TEXT^{+}condition than in the SPEECH^{+}condition type at least once, this difference is not significant.[1] Only the differences between SPEECH^{+}and SPEECH^{-}($\chi^2_{(2)} = 15.73$, $p < .001$) and between TEXT^{+}and TEXT^{-}($\chi^2_{(2)} = 6.05$, $p < .05$) are significant. Finally, the rule of matched modality might apply only when it is not overruled by other relevant factors. Consider table 7, which zooms in on the percentage of typed answers for the four conditions. For all systems it holds that the probability of typing digits is higher than the probability of typing words. For the TEXT^{+}condition this difference is significant ($p = .05$). This (marginal) difference might indicate that if speaking is not obviously more efficient (i.e., the input does not require many keystrokes) and there is no effect of a bad recognizer (which leads subject to type more anyway, see above), then subjects are more likely to match modalities. Closer inspection of the data reveals that only a limited number of subjects is responsible for this finding, which implies at most that for some subjects the rule of matched modality might apply when no other factors do.

[1] $\chi^2_{(2)} = 4.27$. Since this χ^2-value is relatively close to the significance-threshold, two additional subjects were tested in the TEXT^{+}condition: both spoke exclusively. From this we conclude that the difference between SPEECH^{+}and TEXT^{+}is really not significant, and can not be attributed to an insufficient number of subjects.

4. DISCUSSION AND CONCLUSION

A strong influence of speech recognition performance on modality selection was found. When being examined by a system with a realistic recognizer, subjects type significantly more often than they do when being questioned by systems with a perfect recognizer. The majority of the modality switches is related to speech recognition errors. This is in line with the findings of Oviatt and co-workers. Moreover, it was found that numerical answers are somewhat more likely to be typed than other answers. We feel that this is related to efficiency, because the numerical answers could always be entered with a highly limited number of keystrokes. However, on the basis of this experiment we can not rule out the possibility that Oviatt's contrastive functionality was at work. The experiment does provide evidence for the influence of efficiency on modality selection, since subjects generally prefer speech as input modality and indicate that they do so on the basis of speech being faster and more easy to use.

Contrary to our initial expectation, there appears to be no effect of the modality of the system's prompt on subjects' modality selection, as the rule of matched modality suggests. There are two possible explanations for this. The first is that the rule of matched modality, which is presented as an informal observation, is simply wrong. However, a post hoc experiment (described in the appendix) revealed that this is not the case. The other potential explanation is that the effect of matched modality is undone by the strong effects of efficiency and recognition performance. The one piece of evidence which might be attributed, with some good will, to an effect of matched modality hints in this direction. In any case, it is clear that an effect of matching modalities is much weaker than that of recognition performance and efficiency. Moreover, it seems likely that further factors not taken into account here have an influence on modality selection, and these may also overrule matched modality.[2]

APPENDIX: DO PEOPLE MATCH MODALITIES?

The rule of matched modality states that it is polite to respond to people using the same method that they used to contact you. We tested this hypothesis by conducting a small experiment with 26 subjects, all working in our institute. Of these 26 subjects, 12 were given a (fictitious) note from the secretary with the following message: "mrs. X called. It is urgent. Could you contact her as soon as possible?", followed by both phone number and email-address of X (who works elsewhere). The remaining 14 subjects received the following email from one of the authors of this paper: "*Subject*: URGENT: do you have the book by Reeves and Nass? *Body*: Hi Z, Because of a deadline I need to check something today in The Media Equation. However, the book is not in the library. I've heard that you have it. If so, can I borrow it?", followed by both phone number and email-address of said author.

The rule of matched modality predicts that people who were (led to believe that they were) called, call back; and those who receive an e-mail return one. Table 8 shows that people indeed have a strong tendency to match modalities ($\chi^2_{(1)} = 7.54, p < .01$). As a final remark, note that matching modalities is not necessarily the most polite thing to do. One subject who received an e-mail did not send an e-mail back, but came walking into the office actually *bringing* a copy of the requested book.

Table 8: Numbers of subjects who match modalities after receiving a phone-call or an e-mail.

	modalities	
	matched	non-matched
telephone	10	2
e-mail	10	4
total	20	6

Acknowledgements Thanks are due to Ielka van der Sluis for help in carrying out the experiments.

5. REFERENCES

1. Baber, C., *Speech technology in control room systems: A human factors perspective*, Ellis Horwoord, Chichester, UK, 1991.
2. Chalfonte, B., Fish, R., and Kraut, R., Expressive richness: a comparison of speech and text as media for revision, *CHI* '91, 21-26, New Orleans, 1991.
3. Eagly, A. and Wood, W., Inferred sex differences in status as a determinant of gender stereotypes about social influence. *Journal of Personality and Social Psychology* **43**:915-928, 1982.
4. Hirschberg, J. and Whittaker, S., Voicemail overload: the problems of processing voicemail information, manuscript, AT&T Laboratories — Research, 2000.
5. Krahmer, E., Swerts, M., Theune, M., and Weegels, M., The dual of denial: Two uses of disconfirmations in dialogue and their prosodic correlates. *Speech Communication*, to appear, 2000.
6. Oviatt, S., Bernard, J., and Levow, G.A., Linguistic adaptations during spoken and multimodal error resolution. *Language and Speech. Special issue on Prosody and Conversation*, 41 (3-4): 419-422, 1998.
7. Oviatt, S. and Cohen, P., Multimodal interfaces that process what comes naturally, *Communications of the ACM*, **43**(3):45-53, 2000.
8. Oviatt, S. and Olsen, E., Integration themes in multi-modal human-computer interactions. *Proceedings ICSLP'94*, Acoustical Society of Japan, vol 2, 551-554, 1994.
9. Reeves, B. and Nass, C., *The media equation: How people treat computers, television, and new media like real people and places*, CSLI Publications/Cambridge University Press, Stanford, CA, 1996.
10. Shneiderman, B., *Designing the user interface*, third edition, Addison-Wesley, 1998.
11. Swerts, M., Litman, D. and Hirschberg, J., Corrections in spoken dialogue systems, *Proceedings ICSLP 2000*, Peking, China, *these proceedings*, 2000.
12. Weegels, M., Users' (Mis)conceptions of a voice-operated train travel information service, *IPO Annual Progress Report*, Eindhoven, The Netherlands, 1999.
13. Zipf, G.K., *Human behavior and the principle of least effort*, Addison-Wesley, Cambridge, MA, 1949.

[2] For instance, in a study of voice- vs. e-mail, Hirschberg & Whittaker (2000) found that subjects reply to e-mail with a voice-mail (thereby overruling matched modality) if the content is urgent (but see appendix) or when items require immediate attention.