USER EVALUATION OF THE MASK KIOSK*

L. Lamel, S. Bennacef, J.L. Gauvain, H. Dartigues†, J.N. Temem†

LIMSI-CNRS, BP 133, 91403 Orsay cedex, France

†SNCF, Direction de la Recherche et de la Technologie,

45, rue de Londres, 75379 Paris, France

January 26, 2003

Keywords: spoken language systems, speech recognition, speech understanding, natural language understanding, information retrieval dialog

Number of pages: 13

3 tables

4 figures

Published in *Speech Communication*, Vol **38**, (1-2):131-139, Sep 2002.

Abstract

In this paper we report on a series of user trials carried out to assess the performance and usability of the Multimodal Multimedia Service Kiosk (MASK) prototype kiosk. The aim of the ESPRIT MASK project was to pave the way for advanced public service applications with user interfaces employing multimodal, multi-media input and output. The prototype kiosk was developed after analyzing the technological requirements in the context of users performing travel enquiry tasks, in close collaboration with the French Railways (SNCF) and the Ergonomics group at the University College of London (UCL). The time to complete the transaction with the MASK kiosk is reduced by about 30% compared to that required for the standard kiosk, and the success rate is 85% for novices and 94% once familiar with the system. In addition to meeting or exceeding the performance goals set at the project onset in terms of success rate, transaction time, and user satisfaction, the MASK kiosk was judged to be user-friendly and simple to use.

*This work was partially financed by the carried out ESPRIT MASK 9075 project.

Dans cet article nous présentons les résultats de tests auprès d'utilisateurs du kiosque MASK (*"Multimodal Multimedia Service Kiosk"*). Le but du projet Esprit MASK était de développer un kiosque d'information et de distribution avec une interface innovante et conviviale combinant les modalités tactiles et vocales. Le prototype a été développé après une analyse des besoins dans le cadre du transport ferroviaire en collaboration avec la SNCF et le groupe d'ergonomie à University College of London (UCL). Le temps de transaction avec le kiosque MASK est réduit de 30% par rapport aux kiosques existants. Le taux de succès est de 85% pour les utilisateurs novices et de 94% pour ceux qui ont déjà utilisé le système (plus de 3 utilisations). Tous les objectifs fixés par la SNCF au début du projet ont été atteints par le prototype, qu'ils concernent le taux de succès, le temps de transaction, ou la satisfaction des utilisateurs. Le kiosque MASK a été jugé convivial et simple d'emploi.

1 Introduction

The ESPRIT Multimodal Multimedia Service Kiosk (MASK) project has developed a prototype kiosk with an innovative, user-friendly interface, combining tactile and vocal input (Chhor and Salter, 1995; Dartigues et al., 1997; Gauvain et al., 1997; Temem, Lamel, and Gauvain, 1999). The prototype kiosk was developed after analysis of the technological requirements in the context of users and the tasks they perform in carrying out travel enquiries. The kiosk has undergone several rounds of user trials, including a series of Wizard of Oz experiments (Dowellet al., 1995; Fraser and Gilbert, 1991) in the early stages of the user interface design, reported at ICSLP'96 (Life et al., 1996). The work reported here was carried out by LIMSI-CNRS, the SNCF (the French Railways) and the Ergonomics group at UCL (University College London).

The physical design of the prototype kiosk has been changed since that reported in Life et al. (1996), and significant improvements have been made to the user interface. The main improvements concern additional features such as a self-presentation illustrating the use of the kiosk and explaining the different types of transactions available; a more intuitive interface with easy switching between tasks (such as information or ticketing); a facial image of a clerk to let the user know what the system is doing (see Figure 2); and a two-level help facility with fixed time-outs (Bernard, 1997; Bernard and Life, 1997).

In this paper we focus on studies of the user assessment of the MASK prototype, using both objective and subjective performance measures. Iterative evaluations were carried out to validate the

Figure 1: Photo of the MASK kiosk.

software integration and user-interface design. A final set of user assessment trials were carried out in May 1998 with 200 subjects at the St. Lazare train station in Paris. An additional set of performance trials compared different interaction modes: tactile only, vocal only, or combined; as well as trials with the same subjects using the MASK kiosk and the standard automated ticket machines located in train stations.

2 System Description

The MASK prototype kiosk, as shown in Figure 1 was designed by the SNCF in collaboration with LIMSI, particularly for aspects concerning the acoustic signal capture. Two prototypes were built, one to carry out spoken language system development work at LIMSI and the other to carry out the user trials at the St Lazare train station in Paris. Various kiosk designs were considered during the project, including a closed cabin so as to provide better acoustic isolation. An open design was preferered however for security and hygiene reasons. The kiosk has a touch screen for tactile input, loud speakers (the bumps in the side panels) and two microphones, located just above and below

the screen. The kiosk is able to provide train timetable and fare information, and simulated ticket purchases.

When not in use, an animated self-presentation is displayed on the screen, illustrating the use of the kiosk and explaining the different types of transactions available: timetable information, fares and reductions, reservations, help, and payment.

Figure 2 shows a picture of one of the interface screens. The face on the right is the clerk, which lets the user know what the machine is doing (waiting, listening, thinking, talking). The button below the clerk is for *push-to-talk*. The text tells the user to maintain the button pushed (i.e., to keep touching the button) while talking. To provide visual feedback, the button changes color when pressed. The push-to-talk mode[1] was found to be easily accepted by most users, greatly simplifying the speech detection problem. Early in the project a series of Wizard of Oz studies were carried out to assess different user interfaces, and the push-to-talk button was found to be comforting to the users since it provided feedback that the machine was listening (Life et al., 1996). For future prototypes other techniques can be investigated to know when the user is talking to the machine, such as using a camera to detect eye movement or facial orientation. The clerk and push-to-talk icons are always present on the screen.

On the left of this screen is a list of trains satisfying the given constraints. At this point in the transaction, the user can select one of the trains vocally (by referring either to its position in the list or to the time) or by pushing on the button to the right of the desired train. The user can obtain earlier or later trains by asking for them or using the arrows.

The lower part of the screen resembles a train ticket, and summarizes the information known by the system. This part of the screen, displaying information required for ticketing, is always displayed. The items are arranged so that they form a succinct sentence in order to encourage users to speak to the system. This example corresponds to the user query "I want to travel from Paris to Lyon, on Thursday November 20th, leaving around 2 p.m." Incomplete items are marked with a question mark. In this example, the items corresponding to Reduction?, the Number of passengers? and the Class? have not been completed.

Once a user has selected a particular train, the main part of the display is changed to provide detailed information about this train. The example shown in Figure 3 is for a train from Nice to Stras-

[1] Since users are assumed to be unfamiliar with speech recognition technology, specific help messages were designed to explain how to use the push-to-talk. Common errors, such as touching and releasing the button immediately rather than keeping it pressed while talking are detected, and an appropriate help message is played.

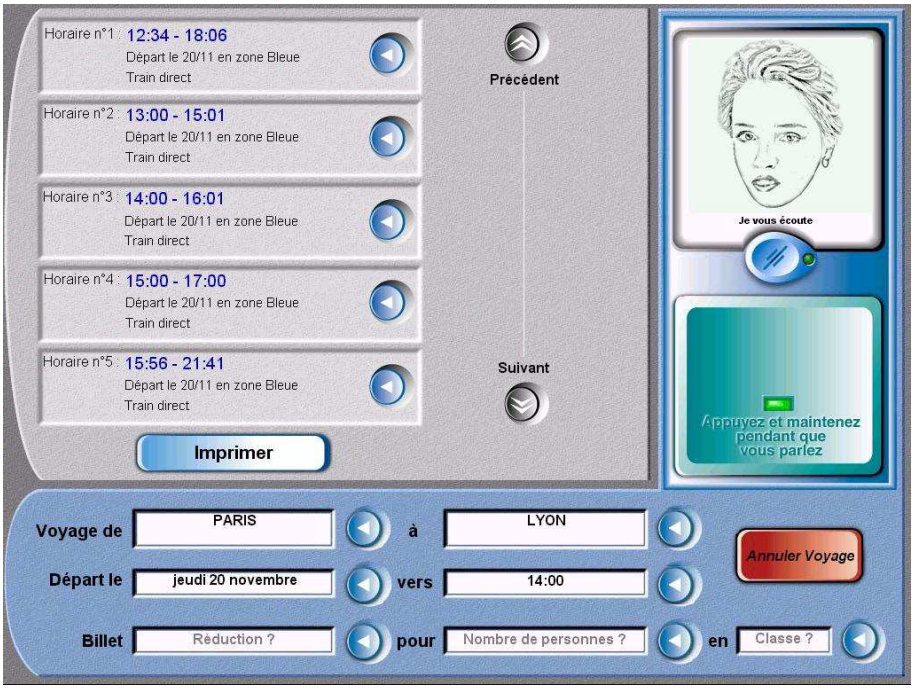

Figure 2: Example user interface from MASK kiosk.

bourg, with a change of trains in Marseille. The user is prompted to specify any missing parameters needed to determine the fare or to make a reservation.

The system architecture is shown in Figure 4. This architecture is a modified version of the LIMSI spoken language system (SLS) (Gauvain et al., 1997), integrating the **Multimedia Interface** and the **Touch Screen**. The main components for spoken language understanding are the speech recognizer, the natural language component consisting of the semantic analyzer and the dialog manager, and an information retrieval component that includes database access and response generation. The speech recognizer (Gauvain et al, 1996; Gauvain and Lamel, 1996) is a medium vocabulary ($\sim$ 2000 words), real-time, speaker-independent, continuous speech recognizer which transforms the acoustic signal into the most probable word sequence. It is a software-only system that runs in real-time on a standard RISC processor. Statistical models are used at the acoustic and word levels. Acoustic modeling makes use of continuous density hidden Markov model (HMM) with Gaussian mixtures. Speaker independence is achieved by using acoustic models which have been trained on speech data from a large number of representative speakers, covering a wide variety of accents and voice qualities. Bigram backoff language models are estimated on the orthographic transcriptions of the training set of spoken queries, with word classes for cities, dates and numbers providing more robust estimates of

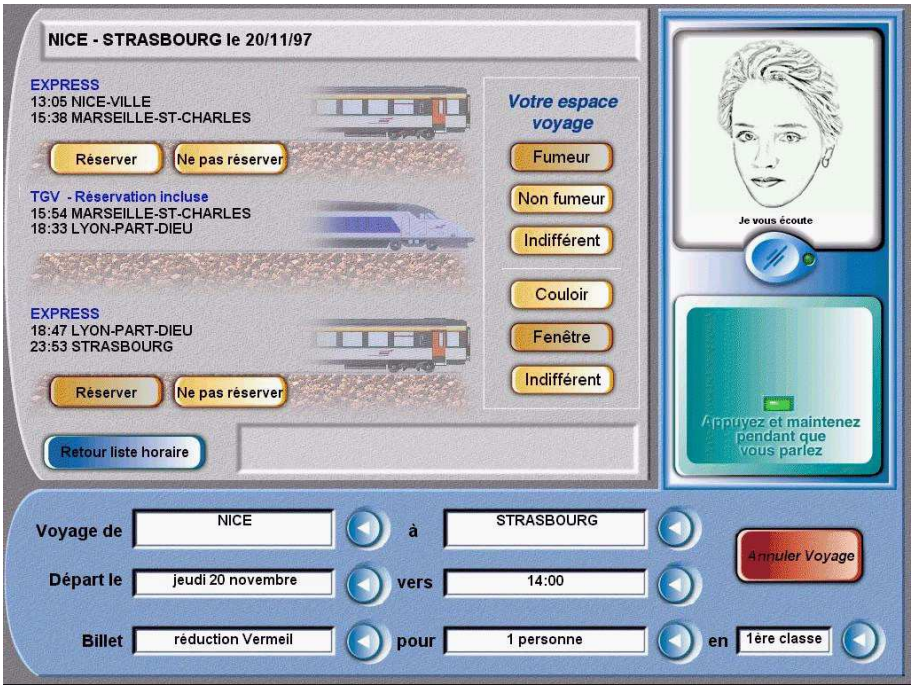

Figure 3: Example screen giving detailed information for a particular train.

the *n*-gram probabilities. The recognition lexicon is represented phonemically and has 1500 entries, including 600 station/city names.

The output of the recognizer is passed to the natural language component which extracts the meaning of the spoken query using a caseframe analysis (Bennacef et al., 1994). Semantic interpretation is carried out in two steps, first a literal understanding of the query, and then its reinterpretation in the context of the ongoing dialog. The major work in developing the understanding component is writing the rules for the caseframe grammar, which includes defining the concepts that are meaningful for the task and their appropriate keywords. The mixed-initiative dialog manager guides the interaction with the user so as to provide the requested information. The dialog manager maintains both the dialog and generation histories, and queries the user so as to obtain information needed for database access. Natural language responses are generated from the dialog state and the information obtained from database access. Information can be returned to the user in the form of synthesized speech or visually, or both. For example, when a list of trains is available it is displayed on the screen and a message is played so as to inform the user. Vocal feedback is provided by concatenation of speech units stored in a dictionary according to the automatically generated response text (Lamel et al., 1993).

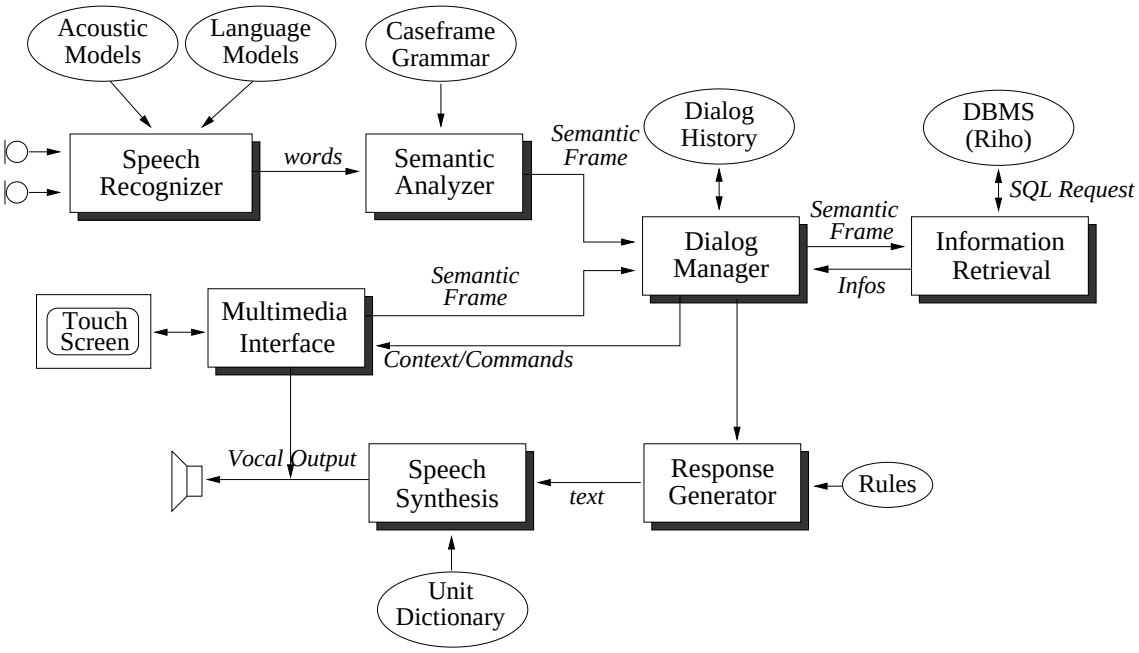

Figure 4: MASK system architecture.

The interaction of the multimedia interface and the spoken language system is via the dialog manager. The multimedia interface interprets tactile commands and generates a Semantic Frame compatible with the SLS. The dialog manager integrates the tactile information into the current dialog context and controls database access. The high level decisions are taken by the dialog manager based on the context and the state of the interface, and low-level presentation decisions are taken directly by the multimedia interface.

An important difference in dialog strategies is offered by the input modes. The tactile strategy is a command driven dialog, where the user must input specific information in order to move on to the next step. Vocal input allows a real mixed-initiative dialog between the user and the system, where the user can guide the interaction or be guided by the system via the help messages. While the two modalities cannot be combined in a single query, a partial specification can be provided with one modality and completed with the other. For example, the user can say “I want a train tomorrow morning”, and then complete the station names by touch. However, these count as separate inputs.

Sex		*Age*				
male	*female*	$18-25$	$26-39$	$40-50$	$51-59$	>60
60	40	23	30	21	13	13

Table 1: Subject population: gender and age.

3 User Trials

Trials with 200 users were carried out to assess the performance of the final version of the prototype kiosk during a 7-day period in April 1998. Complementary user trials were carried out to compare the different input modalities, to compare the MASK kiosk to the current ticket machines, and to assess the effectiveness of the help messages, as well as graphical vs graphical and vocal output.

3.1 Methodology

The user trials were conducted in the St. Lazare train station in Paris. An SNCF hostess selected customers in the train station, and asked if they would be willing to participate in a user evaluation of a new automatic ticket kiosk. Customers that were willing to were escorted to the demonstrator room, located off the main platform. The room characteristics are similar to a passenger waiting area, which is the location envisioned for an initial installation. The hostess selected subjects so as to cover a wide range of ages for each sex, as shown in Table 1. Subjects were given a brief introduction to the purpose of the study and to the tasks to be performed. They were given only a minimal amount of information about the kiosk, such as the possible input and output modalities, but without any specific details. Users were able to learn more about the system capabilities by watching the self-presentation.

100 subjects were divided into 3 subgroups in order to evaluate the kiosk on different tasks: timetable information enquiry (25 subjects), price information enquiry (25 subjects), ticket purchase (50 subjects). In order to assess learning effects, each subject performed the given type of task four times with different scenarios. After each scenario the subject was asked to estimate the time it took to complete it. On completion of the test phase, the user completed a questionnaire and received a 50FF SNCF travel voucher. The questionnaire asked general questions about the subject and their computer experience and travel habits. A series of questions were aimed at their impression of the MASK kiosk such as their overall satisfaction, the ease of use, acceptance of the push-to-talk button, utility of the help messages, and confidence in the vocal input.

Time Task (25)	T1	T2	T3	T4
#inputs	5.2	4.6	3.7	3.2
%speech inputs	23%	27%	46%	56%
$\geq$ 1 spoken action	41%	54%	43%	66%
#help messages	3.9	3.2	2.0	1.2
Transaction time	1’15	0’55	0’43	0’26
Success	79%	70%	97%	99%
Price Task (25)	T1	T2	T3	T4
#inputs	11.4	10.6	9.6	8.7
%speech inputs	16%	20%	25%	25%
$\geq$ 1 spoken action	42%	45%	53%	41%
#help messages	11.0	5.8	3.7	2.8
Transaction time	3’44	2’02	1’46	1’11
Success	96%	89%	98%	99%
Purchase Task (50)	T1	T2	T3	T4
#inputs	13.1	11.9	9.4	9.8
%speech inputs	13%	15%	15%	17%
$\geq$ 1 spoken action	43%	43%	45%	41%
#help messages	9.4	5.8	4.3	2.9
Transaction time	3’26	2’04	1’42	1’35
Success	85%	86%	92%	95%

Table 2: User trial results by task type: time enquiry, price enquiry, and ticket purchase. T1 - T4 correspond to the 1st - 4th time the task was carried out. An input corresponds to the provision of a data item and may be made by touch or speech.

No Difficulty	*Usability*	*Simplicity*	*Satisfaction*
74% (65)	86% (65)	93% (93)	98% (92)

Table 3: User assessment of the MASK kiosk. The project objectives for the user ratings are shown in parentheses ().

3.2 Experimental Results

The main results of the studies are given in Table 2, for the 3 task types: time enquiry, price enquiry, and ticket purchase. Tn shows the averaged results corresponding to the nth transaction of each subject. The first line gives the average number of inputs during the transaction, where an input corresponds to a data item and can be entered vocally or by touch. An input corresponds more or less to a slot in the semantic frame, such as a date, time, station, etc. The second line specifies the percentage of inputs made vocally, and the third is the number of transactions with at least one spoken entry. The fourth line indicates the average number of help messages and the last two lines give the overall transaction times and success rates. It is apparent that the time task is substantially simpler, requiring about half of the actions as are needed for price enquiry and ticket purchase. This is also reflected in the overall transaction times. The effects of subject learning are seen by the reduced number of inputs, help messages, and time as well as an increasing success rate. The percentage of inputs made vocally is around 20%, but for the 4th time enquiry task over half the inputs were spoken. For the time information parts of the price and ticket purchase tasks, a higher percentage of spoken inputs were observed than the task averages (close to those observed for the time enquiry task). On average over 40% of the transactions had at least one spoken input, and for 98% of the spoken inputs a semantic frame was generated. For the remaining inputs no information containing words were recognized, because there were none - the recording contained only a hesitation or noise.

Table 3 shows the overall user ratings compared with the project objectives. 74% of the users never or rarely encountered difficulties in using the system. Subjects were largely satisfied with the usability and simplicity of use, with 98% of them quite or very satisfied.

3.3 Complementary studies

Complementary studies were carried out to determine user preferences between the new kiosk and the automatic ticket machines (APV[2]) currently in service and to assess the role of the different modalities offered by the MASK prototype. Each study involved a new set of subjects, and the subjects were divided in two subgroups, reversing the order of the system or configuration tested.

30 subjects participated in the comparative study of the APV and MASK, carrying out the ticket purchase task. 80% of the subjects preferred MASK, finding it fast and user-friendly, with a 95% preference by people who do not normally use the existing APVs and 75% by those that do. Users preferring the existing APVs had more problems with speech input than users preferring MASK, and being frequent APV users they were able to carry out very efficient transactions. A set of 14 subjects compared a tactile-input only version of MASK to the existing APVs. The MASK transaction success was higher and the user-interface was preferred even though the transactions took longer. This preference was partly due to the user guidance provided via the help messages.

The effectiveness of the help messages was investigated with a set of 15 subjects completing purchasing task without help messages. The help messages were found to be efficient in guiding the user, particularly for the first transaction, and enhanced the subjective evaluation. Many first time users of the MASK kiosk were observed to visually explore the screen for extended periods of time without carrying out any action. Vocal help messages were automatically cued when no user input occurred within the allotted time. The first level messages indicated the goal of the current subtask, and the second level told the user how to specify the required inputs. For the first task (novice users) help messages improved the transaction success rate (85% vs 61%) and led to faster transactions (3'26" vs 4'05"). Subjects also used vocal input more often when help messages were available.

30 subjects compared tactile-only, vocal-only and free mixed modality use of MASK. When forced to choose between one of the two input modalities, speech input was slightly preferred (53%), and had higher subjective ratings and was about 10% faster for the transaction compared to tactile-only. However speech-only was perceived as inefficient if the user needed to repeat, and had a higher average transaction error rate (15% vs 5%). Users preferring touch found it simpler and quicker, and were successful with their tasks. Those preferring speech were less accustomed to the APVs and their preferences were not affected by the success rate. When subjects were allowed to mix modalities, they were able to follow their preferences, reducing the average transaction time and increasing the average success.

[2] Automatique pointe de vente.

4 Discussion and Conclusions

In this paper we have given an overview of the MASK prototype kiosk enabling interaction through the co-ordinated use of multimodal input (speech and touch) and multimedia output (sound, spoken messages, graphics and text). In order to achieve this goal, technical advances were required to allow real-time interpretation of user data entries via multiple input modalities and real-time integration of multimedia feedback to guide the user. A major consideration was the ability to interact effectively with naive users. The help facility was found to efficiently guide users, particularly novices, without disrupting their attention. Since users are not expected to be familiar with speaking to a machine, several help messages describing how to use the push-to-talk button were included.

The user trials demonstrated that for this task multimodality is more efficient (faster and easier) than monomodality as some actions are better carried out by voice and others by touch. Users perceived the relative efficiencies of the input modalities and selected the mode which was most convenient in the particular circumstance. Speech was often used to specify a related set of properties, such as the departure and arrival stations, and the travel date *("I'd like a round-trip ticket from Paris to Lyon today around noon.")*. Touch input was prefered to select a train from a proposed list or to respond to simple questions such as the class of travel and seating preferences (smoking/non-smoking, aisle/window). Spoken input was also found to be more efficient to select amongst a large number of possibilities (e.g., stations) where touch input is laborious. These studies also showed that subjects performed their tasks more efficiently as they became familiarized with the MASK system, learning to exploit the vocal input and benefiting from the multiple modalities.

An average transaction success rate of 93% was obtained on 400 transactions with 100 subjects, while reducing the average time by 30% (from 2'25" with the APV to 1'41" using MASK). The user population was divided roughly in half, with 53% preferring speech because it was faster, simpler and more entertaining than using the touch screen and the remaining 47% preferring touch, mainly due to their familiarity with this modality. Forcing users to use the non-preferred modality led to longer transaction times. Users preferring speech took an average of 1'21" to purchase a ticket only using speech input and 2'14" when required to use only tactile input. (For these subjects, free use of the input modalities had an average transaction time of 1'16"). Users preferring the touch screen took 1'36" to purchase their ticket by touch and 2'3" by speech, compared to 1'32" when free to choose. The free use of multimodalities is seen to provide only a slight reduction in transaction time compared to the preferred input modality as subjects tended to use their preferred input mode. A much larger

reduction in transaction time is observed compared to when subjects were constrained to use their less preferred input mode. The user preferences for spoken or tactile input were observed to be related to their familiarity with the currently available APVs. This distinction can be expected to disappear with continued use of the MASK kiosk, as users learn to optimally perform their transaction. Only with long term studies of the kiosk in free use can the real contribution of the two input modes be assessed.

One concern on the part of the system developers was that users would be hesitant to speak to a kiosk in public. This was not perceived as a problem for the users, as 87% of the subjects said they would be likely to use speech input if the MASK kiosk were located in a train station.

Most subjects preferred the new kiosk design, with a lower preference expressed by frequent users of the current kiosks who are used to carrying out their transactions.

In order to assess usability under operational conditions, the next step, which is still under discussion, is to place a more secure kiosk in the station, so as to provide free access by users needing information or desiring to purchase a ticket. While this will provide very valuable data, without control on the users or tasks, analysing the results will be much more difficult since there will be no indication of transaction success or user satisfaction.

5 Acknowledgements

We gratefully acknowledge the participation of Myriam Vergnes, Sylvianne Lemercier, and Franck Bernard in conducting and analyzing the user trials. This work was carried out in collaboration with the Ergonomics Group at UCL (University College London, 24 Bedford Way, London WC1H0AP, U.K.), a partner of the MASK project. Martin Colbert from UCL carried out the performance analysis on the data obtained in the user trials.

REFERENCES

[1] S.K. Bennacef, H. Bonneau-Maynard, J.L. Gauvain, L. Lamel, W. Minker (1994), "A Spoken Language System For Information Retrieval," *Proc. ICSLP'94*, Yokohama, **3**, pp. 1271-1274, September.

[2] F. Bernard (1997), "Le choix d'un système d'aide vocal pour une borne multimédia et multimodale d'information et de vente dans le domaine ferroviaire," *Proc. Interfaces'97*, pp. 287-289, Montpellier, France.

[3] F. Bernard, A. Life (1997), "The timing of prompts for a multimodal-multimedia public service kiosk," *Proc. IEA'97*, **6**, pp. 264-266, Tempere, Finland.

[4] E. Chhor, I. Salter (1995), "The MASK Project," *Human Comfort & Security Workshop*, Brussels, October.

[5] H. Dartigues, F. Bernard, A. Guidon, J.N. Temem (1997), “The MASK project : new passenger service kiosk technology,” *World Congress on Railway Research ’97*, Florence, pp. 513-518, November.

[6] J. Dowell, W. Lukau, I. Salter, Y. Shmueli, A. Life (1995), “Designing the multimodal speech interface to a public travel facility,” *International Ergonomics Association World Conference 1995*, Rio de Janeiro, October.

[7] N. Fraser, G. Gilbert, “Simulating speech systems,” *Computer Speech & Language*, **5**, 81-99, 1991.

[8] J.L. Gauvain, S. Bennacef, L. Devillers, L. Lamel, R. Rosset (1997), “Spoken Language component of the MASK Kiosk” in K. Varghese, S. Pfleger (Eds.) “Human Comfort and security of information systems”, Springer-Verlag.

[9] J.L. Gauvain, J.J. Gangolf, L. Lamel (1996), “Speech Recognition for an Information Kiosk,” *Proc. ICSLP’96*, Philadelphia, pp. 849–852, October.

[10] J.L. Gauvain, L. Lamel (1996), “Large Vocabulary Continuous Speech Recognition: from Laboratory Systems towards Real-World Applications,” *Institute of Electronics, Information and Communication Engineers*, J79-D-II:2005–2021, December.

[11] L. Lamel, S. Bennacef, J.L. Gauvain, H. Dartigues, J.N. Temem (1998), “User Evaluation of the MASK Kiosk,” *ICSLP’98*, Sydney, pp. 2875-2879, December.

[12] L. Lamel, J.L. Gauvain, B. Prouts, C. Bouhier, R. Boesch (1993), “Generation and Synthesis of Broadcast Messages,” *ESCA-NATO Workshop on Applications of Speech Technology*, Lautrach, Germany, pp. 207-210, September.

[13] A. Life, I. Salter, J.N. Temem, F. Bernard, S. Rosset, S. Bennacef, L. Lamel (1996), “Data Collection for the MASK Kiosk: WOz vs Prototype System,” *Proc. ICSLP’96*, Philadelphia, pp. 1672–1675, October.

[14] M.A. Life, J.B. Long, B.P. Lee (1994), “Providing human factors knowledge to non-specialists: a structured method for the evaluation of future speech interfaces,” *Ergonomics*, **37**, 1801-1842.

[15] MASK Consortium (1998), “Final report of the ESPRIT MASK project,” August.

[16] J.N. Temem, L. Lamel, and J.L. Gauvain (1999), “The MASK, demonstrator: An emerging technology for user-friendly passengers kiosk,” *Proc. World Congress on Railway Research*, Tokyo, October.