Data Collection for the MASK Kiosk: WOz vs Prototype System

A. Life, I. Salter[†]
Ergonomics Unit, UCL, 26, Bedford Way, London, U.K.

J.N. Temem, F. Bernard
SNCF, 45, rue de Londres, Paris, France

S. Rosset, S. Bennacef, L. Lamel
LIMSI-CNRS, Orsay, France

ABSTRACT

The MASK consortium is developing a prototype multimodal multimedia service kiosk for train travel information and reservation, exploiting state-of-the-art speech technology. In this paper we report on our efforts aimed at evaluating alternative user interface designs and at obtaining acoustic and language modeling data for the spoken language component of the overall system. Simulation methods with increasing degrees of complexity have been used to test the interface design, and to experiment with alternate operating modes. The majority of the speech data has been collected using successive versions of the spoken language system, whose capabilities are incrementally expanded after analysis of the most frequent problems encountered by users of the preceding version. The different data requirements of user interface design and speech corpus acquisition are discussed in light of the experience of the MASK project.

1. INTRODUCTION

The aim of the Multimodal-Multimedia Automated Service Kiosk (MASK) project is to develop an advanced service kiosk that will be installed in an SNCF train station in Paris to provide travel information and tickets[2]. The consortium has analysed the technological requirements of the service kiosk in the context of users and the tasks they perform in carrying out travel enquiries. The kiosk should improve the effectiveness of such services by enabling interaction through the coordinated use of multimodal inputs (speech and touch) and multimedia output (sound, video, text, and graphics). Vocal input is managed by a spoken language system, which aims to provide an effective interface between the user and the computer through the use of simple and natural dialogs. The design of the user-dialog has been carried out after analysing human-computer dialogs obtained via Wizard of Oz (WOz) experiments[3] and using preliminary versions of the spoken language system[7, 5].

The development of user interfaces (UIs) often necessitates and iterative development strategies involving simulation and prototyping. While such strategies have some potential drawbacks from the software engineering perspective (e.g. incoherence in design; premature design commitment); these disadvantages are outweighed by the benefits obtained by observing user behavior which is often complex and not well-understood. The MASK UI is novel and complex, and as such is an example of a system for which the iterative design strategy is appropriate and necessary.

SLS development requires a large corpus of training data: the collection of such spoken language data represents a significant portion of the work in developing a system. The use of additional acoustic and language model training data has been shown to almost systematically improve performance in continuous speech recognition[6]. WOz simulation appeared to offer an ideal means of collecting data to meet both of these requirements. However, this paper shows how the UI and SLS needs of the MASK project were found to be different enough that a pragmatic approach using a two stream prototyping strategy was adopted.

2. WIZARD OF OZ SIMULATION OF MASK

The use of people to simulate machines is particularly valuable where design decisions must be made before implemented technology is available and where alternative designs need to be simulated quickly. WOz simulations have been widely used in the speech community[4] to support the early stages of spoken language system (SLS) development, and are increasingly being used in the design of UIs. This technique has been used in the MASK project to empirically support the user interface design early in development. A cyclical development strategy has been adopted, where an initial design based on the specifications of the user and task requirements was implemented and tested.

Three simulation cycles with increasing fidelity have been achieved thus far, each cycle producing a speech corpus[3]. The progressive increase in fidelity is the consequence of two factors: (1) the technical improvements in the quality of the WOz implementation; and (2) refinements of the simulated UI, as a result of the design iterations. Each study sampled the targeted user population, solving a variety of tasks in the train travel domain. In each cycle, a set of scenarios with different complexities was used (see Figure 1).

2.1. Low fidelity WOz simulation

The main objectives of the first simulation (implemented as soon as possible after the specification of the MASK functional requirements) were to collect preliminary speech corpus and to explore some basic constraints for the UI. Experiments were conducted at UCL with 16 computer-literate French-speaking subjects solving tightly constrained travel enquiry tasks in a limited domain (10 stations). Although some critical features of the final interface were

[†] Now at Signes Particuliers, Paris, France

A- *You want to go from Rouen to Bordeaux today. The next train leaves at 11:25 am, but you want to know the time of the following train.*
B- *You want to know the price of a first class round-trip ticket from Cherbourg to Paris. You wish to travel to Paris on Friday, September 22nd, leaving Cherbourg after 6:00pm and to return to Cherbourg on Sunday, September 24th, arriving before 10:00pm.*
C- *You want to travel from Paris to Nantes, arriving before 3:00 pm on Friday, September 1st. You travel in a second class, without a reservation, with a reduction* **Carrissimo**.

Figure 1: Example scenarios used for the sc WOz experiments.

modeled (such as the ability to understand relatively unconstrained speech), the simulation did not accept tactile input, handled only a small set of services (no purchases or printing), and used rudimentary graphics. The study generated data relating to user hesitation and operating errors, providing insight into how users respond to spoken and visual feedback, and patterns of user initiative during the interaction.

931 spoken queries were obtained, containing a total of 5865 words, with 426 vocabulary items. While this data was sufficient to explore the patterns of interaction behavior for the next design iteration, it resulted in a very limited and narrow speech corpus, with little variation in the expressions used by different subjects.

2.2. Medium fidelity WOz simulation

The objectives of the second simulation were to extend the spoken corpus, to examine how users would use a fully integrated multimodal interface, and to study means of optimising interaction efficiency for both novices and experts. A larger subset of the task domain was simulated, enabling users to specify more details of journeys (500 stations, preferred routings, and return travel needs), and to purchase tickets. The scenarios were significantly less constrained than in the low-fidelity simulation, allowing the users more flexibility in problem solving. The UI was multimodal, offering the possibility for interaction using speech and touch. The simulated SLS was able to understand natural language queries. System prompts and help messages were provided by playback of pre-recorded speech.

40 subjects solved an average of 10 scenarios, in a session lasting approximately an hour. The broader domain coverage combined with more general scenarios, resulted in a more variation in query style. Only a portion of this data was transcribed consisting of 891 queries from 18 speakers. These contained a total of 3691 words, with 364 distinct words (excluding word fragments) and an average of 4.1 words per query. The study generated behavioral, timing, and error data for UI design (e.g., integrated use of speech and touch; use of visual displays such as maps and browsers; and changing patterns of interaction with experience). These experiments indicated that visual feedback helped users keep track of the dialog, and that vocal output resulted in slowed performance, as users waited until the end before continuing the transaction. However, the speech corpus obtained was disappointingly small and the overhead in developing this WOz rig was larger than had been anticipated, demonstrating limitations of the simulation approach.

2.3. Preliminary conclusions of WOz experiments

Based on the WOz experiments in London, it was apparent that the simulations were not particularly effective for the dual purpose of corpus collection and ergonomic studiy and evaluation of UIs. Design of the UI benefits from relatively short, frequent simulation development cycles, using a small number of subjects (only enough to ensure that the conclusions are general). Furthermore, at the low level of design (dialog control mechanisms, screen and message design etc.), useful data can be collected in the context of quite limited system functionality, and with a narrow representation of the domain.

In contrast, collection of a representative speech corpus is crucial for SLS development. Therefore it is necessary to ensure good domain and functionality coverage. The low and medium fidelity simulations for MASK did not meet this requirement. Furthermore, the speech corpus must be of sufficient size to represent the population of likely users of the final system, and of their speaking style when operating the system. The number of subjects being tested on the MASK simulations was insufficient to ensure adequate representation.

Because of these differences in data requirements, we switched to a two stream data collection strategy, where the strengths of the WOZ technique for flexible changes in the implementation were exploited for UI design, while the speech corpus was collected in parallel using a prototype SLS.

2.4. High fidelity WOz simulations

The final phase of WOz simulation studies were designed to refine the low level UI design, to empirically test certain critical design options, and to evaluate the performance of the final simulated UI. The simulation operated within a restricted domain, with functionality close to that planned for the final system. Users were able to conduct an extended range of tasks relating to timetable, fare and route enquiries; reservation; and ticket purchases. The UI allowed both unconstrained natural speech and tactile entry, testing issues such as push-to-talk, combined modalities (speech or touch dominant), automatic help messages, and the effects of simulated recognition errors on performance. Written text, animated graphics and video were presented, as well as speech output (consisting of 153 pre-recorded contextual messages by a female speaker).

A total of 13 versions of the UI were implemented and tested over the three-week WOz study period. The experiments were conducted at the Gare St. Lazare in Paris with 110 representative subjects recruited by the SNCF. Audio and video recordings (54h) were made of the sessions and the computer logged and time-stamped all exchanges. At the end of each session, the subject was debriefed and completed a questionnaire.

The **basic UI implementation** was refined over a 3 day period, during which 21 subjects used the system and modifications were made to deal with observed difficulties.

Alternate UI options: Two independent groups of 14 subjects subsequently compared **automatic speech detection** with **push-to-talk**. With animated feedback and error messages to guide users, the push-to-talk mechanism was positively received. This is consistent with previous observations that subjects tend to not mix input modes, preferring either speech or touch, and rarely combine the two within the same utterance.

The effect of different fixed delays of 2, 5 and 10 seconds, and an 'adaptive' delay for providing **promts** and **help messages** to users was studied using 33 subjects divided in 4 groups. The basic assumption is that novice users hesitate when they need guidance, whereas hesitations of expert users are more likely to be due to off-line factors, such as making decisions. A fixed 10s delay was found to be too long and drastically lengthened the transaction time. An adaptive delay appeared to be the most efficient solution for the sample of users tested.

Simulation of speech recognition errors: In designing the MASK UI we have aimed for collaborative achievement of task goals, enabling users to take initiative and to freely switch modalities. If persistent problems arise, the machine should take the initiative to guide the user, if need be encouraging them to revert to the most reliable channel (i.e. touch). The behavior of 12 subjects was studied solving scenarios in which a station name was "misrecognised" on the user's first attempt to enter it (e.g. user said Cherbourg; system recognised and displayed Strasbourg). Users noticed and corrected the errors. Although most corrections were made vocally, some users did switch to the touch modality.

The **final UI simulation** implementing the desirable characteristics identified in the comparative studies was tested with 29 subjects. It included a push-to-talk button; an adaptive delay help prompt system; and a cooperative error recovery strategy. The results analyzed thus far are encouraging and reinforce our confidence that we will be able to achieve the project objectives of reducing the average transaction time and increasing the success rate.

2.5. Summary of WOz simulations

Based on the WOz studies we are able to conclude that such experimentation is effective to assess different UI configurations. In particular, the push-to-talk mode was found to be easily accepted by users, and greatly simplifies the speech detection problem. Given the expected difficulties of automatic speech detection in noisy environments, such capability will not be used in the MASK project. Providing help messages based on adaptive delays (rules accounting for user behavior) affects the nature of the dialog and the complexity of spoken utterances. Speech understanding errors appear to be not too critical to overall performance due to the multimodal-multimedia nature of the system, as users noticed and corrected the errors.

3. PROTOTYPE SYSTEM DATA COLLECTION

The collection of spoken language corpora represents a significant portion of the work in developing a spoken language system. Our experience is that as system performance improves, subjects speak more easily using longer and more varied sentences. They are also more likely to perceive errors as their own fault, rather than the system's. As a result they continue to speak relatively naturally to the system, enabling us to record representative spontaneous speech, which is in turn used to improve the system. In addition to the acoustic and language modeling data needed for training, the subjective evaluation provided by the users helps us to assess progress in system development.

The MASK spoken language system[5] consists of a speaker-independent continuous **speech recognizer**, whose output is passed to a **natural language** (NL) component. The NL component is concerned with understanding the meaning of the spoken query and

A- *You want to go from Tours to Rennes next Monday. You would like to arrive after 6:00 pm.*
B- *You want to go from Limoges to Montpellier, next Friday. Ask for a train with a sleeping car.*
C- *You are travelling from Paris to Orléans tomorrow. You have a reduction* **Modulopass**. *Reserve an aisle seat in a second class, smoking car.*

Figure 2: Example scenarios used for data collection with the SLS system.

includes the **semantic analysis**[1] and **dialog management**. Natural language responses are automatically generated from the semantic frame, the dialog history and retrieved DBMS information, and synthesized using concatenated speech from stored dictionary units. The vocal feedback is provided along with visual information.

At LIMSI we have recorded over 13,000 queries from 194 subjects, with about 1400 vocabulary items (not including word fragments) found in the 103k words[7]. During the WOz experiements in situ at the Gare St. Lazare in Paris, data was collected with an intermediary version the SLS from 121 subjects recruited by the SNCF. The average duration of the recording session was 30 minutes. We used a set of 12 scenarios with different complexities. Each subject solved 2-4 different scenarios. Three examples of scenarios are given in Figure 2. A total of 3412 queries were obtained containing 28.9k words, of which 1554 were distinct.

Objective evaluation. The objective data concern the total scenario duration and the total number of turns per scenario. The average number of turns is 9 for total duration about 4 min.

83% of the 368 scenarios were successfully solved, with the subject in obtaining the correct timetable information. These results reflect the performance of a system still under development, we can expect much better performance in the final implementation. Although there was an error on 42% of the queries, these errors do not necessarily result in dialog failure (17%). A more detailed evaluation was carried out to assess the recognition and understanding performance for slots relevant for semantic analysis. Table 1 shows the percentage of slots not recognized/understood for the **departure-city**, **arrival-city**, **departure-time**, **arrival-time** and **departure-date**. The error rates correspond to the number of erroneous slots divided by the total number of slots for each type.

Error	*dep-city*	*arr-city*	*dep-time*	*arr-time*	*dep-date*
#slots	78	80	216	95	86
Reco	5.2%	4.2%	18.5%	8.2%	29.6%
Und	3.6%	4.4%	7.0%	0.5%	6.0%

Table 1: Recognition and understanding error rates on semantic slots.

To evaluate the understanding performance, we distinguish errors due to recognition errors from those due to understanding. Table 2 shows the recognition and understanding error rates averaged across all queries. The query understanding error rate is seen to be about 1/3 that of the recognition error rate, because not all recognition errors lead to an error in understanding.

To evaluate the dialog performance, we distinguish errors due to the understanding component, to the dialog manager and to database access. As shown in Table 3, the vast majority of dialog errors are due to recognition and understanding errors, and very few are attributed to the dialog management.

Recognition	*Understanding*
16.2%	5.4%

Table 2: Average recognition and understanding error rates by query.

#Responses	*Correct*	*Reco+Und*	*Dialog*	*DB access*
3412	57.9%	35.7%	5.4%	1.0%

Table 3: Source of dialog errors measured by response.

Subjective User evaluation: Qualitative measures such as the satisfaction of the user are necessary to assess the performance of spoken language systems in real-word applications. In order to assess the overall system performance, subjects complete a questionnaire addressing the ease-of-use, reliability, and satisfaction with the MASK system. The responses of 104 subjects[1] have been analyzed. The overall results are shown in Table 4 on a scale of 10. Subjects were asked how often they travel by train, how they obtain their ticket, and about their computer experience. They also were asked to specify the good aspects of the system, how it should be improved, and if they would use such a potential system.

Ease-of-use	*Reliability*	*Satisfaction*
7.7	7.0	6.7

Table 4: Overall subjective evaluation

We have also analyzed the data from subjects classied by their observed comfort with the system, age, and travel habits[5]. While "comfortable" users found the system easier to use than novices (difficulty speaking with the system or using the computer), they rated it less reliable and user-friendly than novices. Novices were more likely to doubt the reliability of information obtained from the system, whereas the comfortables criticized problems in understanding or dialog. There was a clear tendancy of younger subjects to assess the system more favorably than the older subjects. This is likely to be correlated with a larger familiarity of younger subjects with computers and service kiosks. Frequent train travelers were slightly more sceptical and dissatisfied with the system than infrequent travelers.[2] In general, users express an interest in using such types of systems, and volunteer to participate in future experiments. 93% (97/104) subjects considered themselves potential users of a future MASK kiosk.

4. DISCUSSION

At the outset of the MASK project we envisioned that simulation techniques could easily serve the dual purpose of prototyping user interfaces and collecting natural spoken language data. A series of WOz simulations were planned to investigate different design possibilities. The experiments were to have increasing fidelity with respect to the UI of the final MASK prototype kiosk. The experience gained in the early (low and medium fidelity) simulations indicated that the needs for UI design and for speech corpus acquisition could not be easily met using the same simulations. UI design needs to have quick feedback on a relatively small, but representative sample of the user population (N=10-20). The simulations need to be as realistic as possible so as to evaluate user behavoir using M4 UIs. In contrast, spoken language understanding system development requires large amounts of data from many different speakers. The spoken data collected testing UI designs is largely insufficient for the SLS development needs.

The two stream strategy described here avoids conflict in data requirements for UI design and corpus development, but at some costs. It is necessary to develop two data collection systems: a working prototype SLS for data collection and different simulators for UI design. Although the data collection SLS will not need to change as frequently as the simulations, it must be updated at regular intervals on the basis of the UI simulation, to ensure that it remains representative of the final system. With regard to the collecting data with the SLS prototype, our experience is that a tightly coupled data collection, analysis, and system revision cycle is needed to obtain representative data.

Positive aspects of the two stream approach are that desired UI functionality can be implemented in the prototype SLS to further validate the interface in a more realistic situation, and the overall system performance can be assessed relatively early in the design process. The experience of the MASK project suggests that the two methods are complementary, and that the additional costs are well worth bearing for the increased quality of data achievable.

[1] While a total of 121 subjects were recorded, 17 were considered trial subjects while we were setting up the system and did not complete a questionnaire.

[2] This may be because frequent travelers are more aware of the fares for common routes, and the fares calculated by the data collection system are estimates based on old information.

5. ACKNOWLEDGEMENTS

The research described here was funded by the Eurpoean Union under the ESPRIT programme (Project no. 9075). MASK is a collaborative project of UCL, LIMSI-CNRS, SNCF, MORS and Signes Particuliers. The authors acknowledge the critical contributions of the following in conducting the studies described here. At LIMSI: Christophe d'Alessandro, Laurence Devillers, Boris Doval, Saliha Foukia, Jean-Luc Gauvain, Jean-Jacques Gangolf. At UCL: John Dowell, William Lukau, Yael Shmueli. At SNCF: Herve Dartigues, Alain Guidon, Sylvie Letienne.

REFERENCES

[1] S.K. Bennacef, H. Bonneau-Maynard, J.L. Gauvain, L. Lamel, W. Minker, " A Spoken Language System For Information Retrieval," *Proc. ICSLP'94*.

[2] E. Chhor, I. Salter, "The MASK Project," presented *Human Comfort & Security Wshop*, Brussels, Oct. 1995. (to appear).

[3] J. Dowell, W. Lukau, I. Salter, Y. Shmueli, A. Life, "Designing the multimodal speech interface to a public travel facility," *International Ergonomics Association World Conference 1995*, Rio de Janeiro, Oct. 1995.

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

[5] J.L. Gauvain, S. Bennacef, L. Devillers, L. Lamel, S. Rosset, "The Spoken Language Component of the MASK Kiosk," presented *Human Comfort & Security Wshop*, Brussels, Oct. 1995. (to appear).

[6] J.L. Gauvain, L.F. Lamel, G. Adda, M. Adda-Decker, "The LIMSI Continuous Speech Dictation System: Evaluation on the ARPA Wall Street Journal Task," *IEEE ICASSP-94*.

[7] L. Lamel, S. Rosset, S. Bennacef, H. Bonneau-Maynard, L. Devillers, J.L. Gauvain, "Development of Spoken Language Corpora for Travel Information," *Eurospeech'95*.