論文 / 著書情報
Article / Book Information

論題(和文)	User Interface Evaluations for a Multimodal ASR-Driven Train Timetables Application
Title(English)	
著者(和文)	Josef R. Novak, Sadaoki Furui
Authors(English)	JOSEF R NOVAK, SADAOKI FURUI
出典(和文)	日本音響学会2010年春季講演論文集, , No. 1-Q-3, pp. 209-210
Citation(English)	, , No. 1-Q-3, pp. 209-210
発行日 / Pub. date	2010, 3

User Interface Evaluations for a Multimodal ASR-Driven Train Timetables Application*

◎ Josef R. Novak Sadaoki Furui
Tokyo Institute of Technology, Tokyo, Japan
{novakj,furui}@furui.cs.titech.ac.jp

1 Introduction

User Interface (UI) evaluations are a key requirement to developing successful applications, both in the course of research pursuits, and pursuant to the development of commercial applications. Voice-driven UIs (VUIs) present special challenges in that additional dimensions such as cognitive load and linguistic flexibility must be carefully considered so as to minimize complexity while simultaneously maximizing the potential user base and usage contexts. Similarly, Multimodal User Interfaces (MMUIs) such as the iPhone and other recent smartphones present their own set of design challenges.

In this paper we report the results of a simple VUI comparison involving three competing interfaces designed for an iPhone-based train-timetables application. As in our past work such as [2], we focus on the Question-Answering (QA) approach to voice search, which emphasizes the higher information content of precise requests, and the greater naturalness that such queries represent compared to spoken keywords. In particular we look at the relative merits of a system that supports QA style natural language input, a system which supports only keyword queries, and a combined interface which emphasizes natural language queries but also supports keyword-based queries as an ancillary back-off.

Test subjects were asked to evaluate each of the interfaces using criteria recommended by ELRA [3], and to rate various aspects of the design and interaction experience on a 5-point Likert scale. In addition to subjective user-based evaluations, relevant objective aspects of the system were also evaluated including average time-to-task, and task completion rate.

Throughout the experiments, an iPhone3G was employed as the client, and a distributed client-server system based on the T^3 decoder [1] was utilized for all Automatic Speech Recognition (ASR) needs.

2 Voice and Multimodal UI Concerns

The primary goals of MMUIs and VUIs are two-fold: to increase the naturalness of human-computer interactions, and to improve the general robustness of computer driven applications. In the latter case, particularly where MMUIs are concerned, this is achieved by providing multiple, alternative and complementary methods for acheiving application goals, thereby facilitating error recovery. Furthermore, voice driven MMUIs such as those implied by the iPhone differ significantly from other typical VUIs such as automated call-centers, or telephone-based dialogue systems. In contrast to traditional systems, which tend to impose a heavy cognitive load on users, MMUI environments promise to significantly reduce cognitive load through clever use of the visual display and haptic capabilities of the screen. Nevertheless the increased versatility that the growing plethora of these devices provide does come at a price. Specifically it puts a new burden on developers and researchers to avoid abusing these features. Some of the more significant concerns that influenced the design of the present interface include,

- Avoiding presenting information in competing modalities,
- Maximizing the potential individual benefit of each input/output modality, in particular the speech and touch elements, so as to minimize cognitive load,
- Providing natural, obvious visual cues, and
- Ensuring that alternative input methods to speech are also provisioned to be used when environmental factors, usage issues, or cultural concerns require them.

These are described in greater detail in [4].

3 Task Description

The evaluation task consisted of two major components. The first was custom web application which supplied a random selection of keywords suitable for formulating queries for train timetables searches. The set of keywords for a given query always included a departure station and an arrival station. These stations names were drawn randomly from the full collection of 10,000 rail and subway stations. Other optional keywords including times, departure, arrival, last station, first station were presented for a fraction of the prompts. Approximately 50% of the presented queries contained only station names, while the remaining 50% of the time one of last station, first station, or a random time were presented with equal likelihood. In the case where a time was presented one further keyword, “departure” or “arrival” or nothing were presented with equal probability. In addition to the query keywords the web application presented a google maps overlay that was tied to the query keywords in order to link them together into a route, with the intent to further encourage spontaneity in query formulation.

The second component was the iPhone client, for which three separate interfaces were prepared. The interfaces can be seen in Figures 1 and 2. The first two interfaces emphasize a single input approach, the first being a natural language input scenario and the second emphasizing keyword-based queries. The third interface combines both modalities, and while it emphasizes natural language input, keyword-based voice input is permitted.

The experiment consisted in presenting each test subject with a series of 10 separate query prompts for each of the three different interfaces. The order in which the interfaces was presented was varied for each subject. After each evaluation, the subject was asked to fill out a short questionaire which asked the subject to evaluate the following criteria using

*マルチモーダル音声認識駆動乗換案内アプリのユーザー・インターフェース評価
ノバック・ジョセフ、古井貞熙

a 5-point Likert scale: overall satisfaction, perceived first interaction success rate, perceived overall system accuracy, ease of use, interface layout, and ease of making corrections.

Figure 1: The natural language interface on the left, and the keyword interface on the right.

Figure 2: The combined interface. Permits both natural language queries, and keyword-based requests.

4 Experimental results

The test group consisted of a total of 8 individuals including 3 females and 5 males. The results from the subjective assessment can be seen in Table 1.

	UI1	UI2	UI3
Overall satisfaction	4.0	3.5	4.2
Perceived 1st success	3.8	3.0	4.3
Perceived overall accy	3.7	3.3	4.2
Ease of use	3.5	3.1	3.7
Interface layout	4.0	3.5	4.2
Ease of corr	3.5	3.0	4.2
Average	3.8	3.2	4.1

Table 1: Average Likert scores for each of the subjective survey questions, corresponding to each of the UI options.

These results indicate that users preferred UI3 on average. In addition to the Likert assessment scores, average time-to-task, task completion rates, and abandonment rates were calculated for each of the interfaces. The results for these objective metrics can be seen in Table 2.

	UI1	UI2	UI3
Avg. time-to-task (s)	21.7	26.8	22.4
Task completion rate (%)	94.3	92.9	94.1
Task abandonment rate (%)	0.5	1.0	0.6

Table 2: Scores for the objective metrics corresponding to each UI.

The average time-to-task refers to the time required for the user to make the input screen match the keyword prompts. Task completion rate refers to accuracy on the first query however, this also includes the N-best lists in order to compensate for the large number of homonyms in the vocabulary. The task abandonment rate refers to the percentage of queries for which the user resorted to text input as a consequence of ASR mistakes. The scores for the objective metrics indicate that UI2 had the longest time-to-task, which is not surprising as it required a minimum of two voice queries, and required further text input for time information. The task-completion rates for UI1 and UI3 are similar. This follows from the fact that they both share the same primary input method. The results for UI2 were again the worst, and investigation of the errors for each of the interfaces indicated that this was primarily to do difficulties with very short station names. In the case of UI1 and UI3 the QA paradigm may provide further contextual information which can be leveraged in these cases. Some users required multiple voice queries in order to obtain the correct results for certain keyword sets, however the task abandonment rates were quite low.

5 Conclusion and Future Work

In this paper we have taken a look at three different Voice driven MMUIs designed for a Japanese train timetables application. We evaluated each of the interfaces based on several well-known, subjective criteria based on a 5-point Likert scale. We also looked at three objective metrics which we believe are relevant to the further development and improvement of our system. The results of the subjective evaluations indicated that the test subjects preferred the QA style interfaces over the keyword-only interfaces. Furthermore, the objective metrics reinforced this view, indicating that for each of the criteria, the QA style interfaces were more efficient as well as marginally more accurate. In future we would like to perform a larger evaluation with more test subjects to confirm the results of this smaller study. We would also like to look at the impact of button arrangement, button size, and the utility of different recording strategies.

References

[1] P. R. Dixon, D. A. Caseiro, T. Oonishi and S. Furui *The Titech Large Vocabulary WFST Speech Recognition System*. Proc. ASRU, pp. 1301–1304, 2007.

[2] E. W. D. Whittaker, P. Dixon, J. Novak and S. Furui *A Prototype Spoken Natural Language Interface for Information Access on Mobile Phones*. Proc. ASJ, 2-3-21, pp. 103–104, 2007.

[3] J. Larson *Ten Criteria for Measuring Effective Voice User Interfaces*. http://www.hlt-evaluation.org/article.php3?id_article=45, 2005.

[4] L. Reeves, J. Lai, J. Larson *Guidelines for Multimodal User Interface Design*. Communications of the ACM, pp. 57–59, 2004.