Empirical Evaluation of Multimodal Input Interactions

Sanjay Ghosh[1,2], Anirudha Joshi[2], and Sanjay Tripathi[1]

[1] Industrial Software Systems, ABB Corporate Research, Bangalore, India
[2] Indian Institute of Technology, Bombay, Mumbai, India
{sanjay.ghosh,sanjay.tripathi}@in.abb.com,
anirudha@iitb.ac.in

Abstract. With variety of interaction technologies like speech, pen, touch, hand or body gestures, eye gaze, etc., being now available for users, it is a challenge to design optimal and effective multimodal combinations for specific tasks. For designing that, it is important to understand how these modalities can be combined and used in a coordinated manner. We performed an experimental evaluation of combinations of different multimodal inputs, such as keyboard, speech and touch with pen etc, in an attempt to investigate, which combinations are efficient for diverse needs of the users. In our study, multimodal combination of three modalities was found to be more effective in terms of performance, accuracy and user experience than that of two modalities. Further, we also inferred the roles that each of the modalities play in a multimodal combination to achieve the usability goals.

Keywords: Multimodal interaction, modality combinations, usability testing.

1 Introduction

The choice of a multimodal interaction depends on its interaction capabilities, the nature of task, application context and users [1]. With a majority of these modern interaction technologies achieving a level of maturity for reaching out to the mainstream mobile and computer applications, what would be the role of traditional interaction methods like mouse and keyboard in the near future? In this study the following research questions were under exploration –

- Which of the multimodal input combinations are efficient?
- Which modality contributes to which of the key usability goals?
- For multimodal combinations, are the usability parameters correlated?

Usability evaluation seems to be the logical method to investigate such questions [1]. In a multimodal setup, use of one interaction technology is also influenced by the co-presence of the other interaction technologies. For such multimodal combinations, "the total usability thus obtained is greater than the usability of each individual modality" [2]. Therefore, for evaluation, combinations of multimodal interactions must consider the as a whole, and not the sum of individual interactions. One of the

S. Yamamoto (Ed.): HIMI/HCII 2013, Part I, LNCS 8016, pp. 37–47, 2013.

objectives of this work is, to evaluate the multimodal interaction, especially for the commonly performed computer tasks. We considered two independent task sets, navigation and editing, to test users' performance. The evaluation was centered on measurement of *performance*, *accuracy* and *user experience* through the following four multimodal combinations; *K+S*, *K+T*, *S+T* and *K+S+T*, where *K* stands for a keyboard, S stands for speech input and T stands for touch input with a pen or finger.

2 Related Works

Broad categories work related to the usability evaluation of multimodal interaction includes, multimodal evaluations through user questionnaires [3], user performance logs [4] [5], both performance logs along with questionnaires [6], Wizard of Oz technique [7] [8], eye tracking [9], model based formal verification methods with the use of Petri-nets [10] or Finite State Machines [11], etc. Ren et al. [5] reported empirical evaluation of Mouse, Keyboard, Speech and Pen for a prototype map and CAD applications. Metze et al. [6] used the post experiment user questionnaires to compare touch and speech modalities independently and their multimodal combination, for a wall mounted GUI based room management system. Similarly, Kaster et al. [12] evaluated the performance of the uni-modal combinations (only a single modality) and bi-modal combinations (two modalities) of mouse with speech, and touch with speech. Wechsung et al. [3] investigated the direct relationship between the bi-modal combinations and uni-modal interactions in terms of user experience based on user's rating on questionnaires. In contrast to these works, we considered only bi-modal combinations and tri-modal combinations (three modalities), because uni-modal interaction may be considered to be just a hypothetical situation without any practical use. Bernhaupt et al. [10] evaluated two mouse and speech, on an industry grade safety critical system by adopting the eye-tracking technique.

Almost all the earlier works did perform usability experiments on their custom developed prototype applications except for Beelders et al. [13], which used Microsoft Word to evaluate speech and eye gaze interactions as a replacement for the conventional typing. In the similar lines, our intent to use commonly performed tasks on computers for our experiments was, to evaluate multimodal combinations catering to a diverse group of users. Thus, our work is unique in terms of the multimodal combinations being evaluated and also the kind of tasks used for user experiments.

3 Method and Experiment Design

Experiments were conducted on IBM ThinkPad X230T, a touch screen enabled tablet computer, with an external keyboard attached. The users were allowed to use the handwriting recognition tool and Windows 'on screen keyboard'. For speech interaction, Microsoft Speech Application Programming Interface [14] was used. A good quality collar microphone was used to capture user's voice. The experiments were performed with ten participants within the age group of 25 to 30 years who were conversant in the use of computers. User recruitment was done using convenient

sampling. Each participant was given speech training and a practice session of 3 hours. The experiment consisted of two categories of tasks, navigation and editing on Windows computer. The goal assigned for a navigation task was to navigate across the Windows help documentation to search for some information and to perform a calculation on the calculator tool. The goal assigned for the editing task was to document few sentences on the WordPad application, and then edit few words out of the text. Each user had to perform four such tasks using four different multimodal combinations as mentioned earlier. Participants were asked to work as natural as possible with the goal to complete the task quickly, with least errors.

Dybkjaer et al. [1] suggested the use of three usability parameters recommended by the ISO for such evaluations namely, efficiency, effectiveness, and user satisfaction. In our study, we renamed those as performance, accuracy and experience. Performance and accuracy were operationalized by the total time to complete a task and the number of errors committed, respectively. The objective was to evaluate the rankings of multimodal combinations in terms of their effectiveness for different tasks.

4 Results and Discussion

Statistical analysis was performed on low level captured data of user's events (screen captures, mouse clicks, pen movements and voice commands). We here present the results of the statistical analysis and our inferences on each of the research questions.

4.1 Which of Multimodal Combinations Are More Efficient?

We made a comparison among all the combinations of input modalities. The objective was to come out with the rankings of multimodal combinations in terms of their effectiveness for different tasks. Also it was investigated whether a combination of three modalities (tri-modal) is more effective than that of two modalities (bi-modal).

Performance. We observed the amount of time taken by the participants in completing the navigation and editing task using all the multimodal combinations one by one. Fig.1 shows the mean and standard deviation of the task completion times.

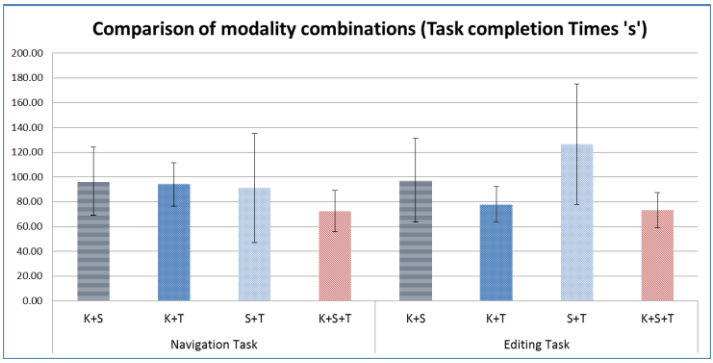

Fig. 1. Mean and S.D. of task completion times for all multimodal combinations

One way ANOVA was applied to analyze the performance variation among the four multimodal combinations. We found a significant variation among the multimodal combinations K+S, K+T, S+T and K+S+T w.r.t. the task completion times for editing task ($F(3, 36) = 5.949$, $p = 0.002$), while this variation was not statistically significant for navigation task ($F(3, 36) = 1.454$, $p = 0.243$). Table 1 shows the result of paired sample t-tests comparing the multimodal combinations performance. For navigation task, K+S+T combination is significantly faster than K+S and K+T; for editing task, K+S+T and K+T combinations are significantly faster than K+S and S+T. All statistically significant values ($p<\alpha$) are indicated by '*'. Additionally, by using t-test we compared the task completion times of bi-modal combinations (K+S, K+T and S+T) and the tri-modal combination (K+S+T). In terms of performance, the tri-modal combination was significantly better than the bi-modal for both navigation ($t(29) = 2.786$, 2-tailed test, $p = 0.009$) and editing tasks ($t(38) = 3.215$, 2-tailed test, $p = 0.003$).

Table 1. T-test results for multimodal comparison w.r.t. the task completion times

	Navigation Task			Editing Task		
Multimodal pairs	df	t stat	sig. '*p*'(2-tailed)	df	t stat	sig. '*p*'(2-tailed)
K+S vs. K+T	9	0.191	0.853	9	2.439	0.037*
K+S vs. S+T	9	0.476	0.645	9	-3.483	0.007*
K+S vs. K+S+T	9	3.586	0.006*	9	3.103	0.013*
K+T vs. S+T	9	0.194	0.850	9	-4.090	0.003*
K+T vs. K+S+T	9	2.867	0.019*	9	0.968	0.358
S+T vs. K+S+T	9	1.435	0.185	9	3.999	0.003*

Accuracy. We counted the number of errors committed by the participants, while performing the navigation and editing tasks using all the multimodal combinations one by one. Fig.2 shows the mean and standard deviation of the number of errors. One way ANOVA showed that, there is significant variation among the multimodal combinations K+S, K+T, S+T and K+S+T for navigation task ($F(3, 36) = 5.594$, $p = 0.003$), as well as editing task ($F(3, 36) = 8.008$, $p = 0.000$).

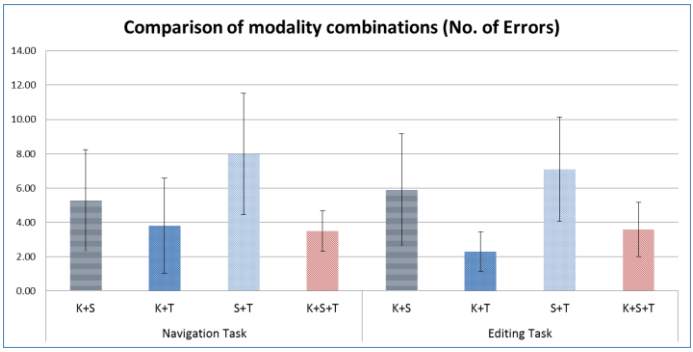

Fig. 2. Mean and S.D. of number of errors for all multimodal combinations

Further, table 2 shows the result of paired sample t-tests comparing the accuracy of the four multimodal combinations for navigation and editing task. Results of the t-tests showed that, for navigation task, S+T combination is significantly less accurate than all other combinations; for editing task, K+S+T and K+T combinations are significantly more accurate than K+S and S+T. Additionally, using t-test we also compared the errors committed by the participants using all the bi-modal combinations (K+S, K+T and S+T) as well as the tri-modal combination (K+S+T). We found that in terms of accuracy, the tri-modal combination is significantly better than the bi-modal combination only for navigation tasks ($t(38) = 2.989$, 2-tailed test, $p = 0.005$). In case of editing task this difference is not statistically significant ($t(33) = 1.919$, 2-tailed test, $p = 0.064$).

Table 2. T-test results for multimodal comparison w.r.t. the number of errors committed

	Navigation Task			Editing Task		
Multimodal pairs	df	t stat	sig. '*p*'(2-tailed)	df	t stat	sig. '*p*'(2-tailed)
K+S vs. K+T	9	1.649	0.134	9	3.592	0.006*
K+S vs. S+T	9	-3.304	0.009*	9	-0.937	0.373
K+S vs. K+S+T	9	2.529	0.032*	9	2.438	0.037*
K+T vs. S+T	9	-4.075	0.003*	9	-4.657	0.001*
K+T vs. K+S+T	9	0.331	0.748	9	-1.709	0.122
S+T vs. K+S+T	9	4.538	0.001*	9	3.312	0.009*

User Experience Level. We analyzed the user experience grades given by the participants using all the multimodal combinations one by one. Fig.3 shows the mean and standard deviation of the user experience levels.

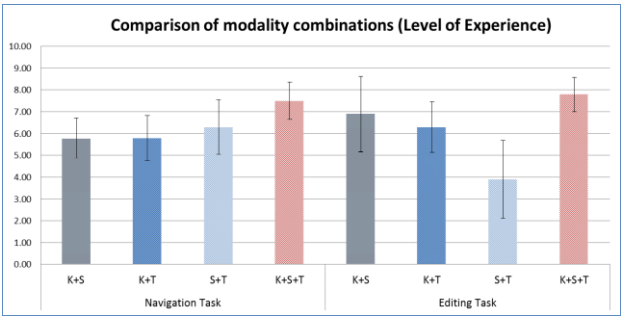

Fig. 3. Mean and S.D. of number of errors for all multimodal combinations

One way ANOVA showed that, there is significant variation w.r.t. the user experience levels among the multimodal combinations K+S, K+T, S+T and K+S+T for navigation task ($F(3, 36) = 6.127$, $p = 0.002$), as well as editing task ($F(3, 36) = 13.629$, $p = 0.000$). Further, table 3 shows the result of paired sample t-tests comparing the user experience level of the four multimodal combinations. Results of the

t-tests shows that, for navigation task, the K+S+T combination has significantly higher level of user experience than all other combinations; for editing task, S+T combination has significantly lower level of user experience than all other combinations. Additionally, using t-test we also compared the user experience levels for bi-modal combinations (K+S, K+T and S+T) as well as the tri-modal combination (K+S+T). We found that in terms of user experience, the tri-modal combination is significantly better than the bi-modal combination for navigation ($t(19) = -4.6208$, 2-tailed test, $p = 0.0002$) as well as for editing tasks ($t(37) = -4.7170$, 2-tailed test, $p = 0$).

Table 3. T-test results for multimodal comparison w.r.t. the user experience level

	Navigation Task			Editing Task		
Multimodal pairs	df	t stat	sig. '*p*'(2-tailed)	df	t stat	sig. '*p*'(2-tailed)
K+S vs. K+T	9	0	1	9	0.818	0.434
K+S vs. S+T	9	-1.103	0.299	9	5.196	0.001*
K+S vs. K+S+T	9	-5.667	0.000*	9	-1.868	0.095
K+T vs. S+T	9	-0.921	0.381	9	3.145	0.012*
K+T vs. K+S+T	9	-3.431	0.748	9	-6.263	0.000*
S+T vs. K+S+T	9	-3.674	0.005*	9	-6.263	0.000*

Table 4 presents the summary of our analysis on effectiveness of different multimodal combinations. We found that the tri-modal combination is more effective than the bi-modal combination. For a multimodal combination to be effective, the involved modalities should complement each other [4].

Table 4. Comparison chart for different multimodal combinations

Task type	Parameter	Multimodal combination ranking
Navigation	Performance	K+S+T > S+T > K+T > K+S
	Accuracy	K+S+T > K+T > K+S > S+T
	User Experience	K+S+T >> S+T > K+S = K+T
Editing	Performance	K+S+T > K+T > K+S >> S+T
	Accuracy	K+T > K+S+T >> K+S > S+T
	User Experience	K+S+T > K+S > K+T >> S+T

where, '>' represents – greater than and '>>' represents – significantly greater than.

4.2 What Are the Roles of Each Modality in a Multimodal Combination?

Here we made a comparison between multimodal combination pairs where, in one of the combination an input modality was present and in the other pair it was absent. This comparison gave us an idea about the role of each input modality towards usability goals, such as, performance, accuracy and user experience level.

Performance. Fig.4 shows the mean and standard deviation of task completion times, for the following multimodal combinations -

- With keyboard (K = K+S, K+T) and without keyboard (No K = S+T)
- With speech (S = K+S, S+T) and without speech (No S = K+T)
- With touch (T = K+T, S+T) and without touch (No T = K+S)

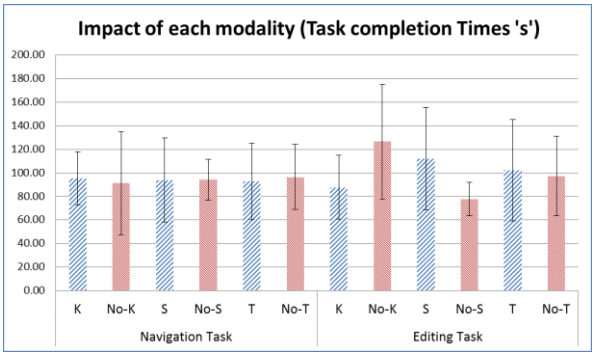

Fig. 4. Mean and S.D. w.r.t. task completion time with/without keyboard, speech and touch

Table 5 shows the result of t-tests comparing the performance of these multimodal combinations. Results shows that, for editing tasks, availability of keyboard significantly increases and that of speech significantly decreases the performance.

Table 5. T-test results of task completion times with and without keyboard, speech and touch

	Navigation Task			Editing Task		
Multimodal pairs	df	t stat	sig. '*p*'(2-tailed)	df	t stat	sig. '*p*'(2-tailed)
K vs. No K	11	0.281	0.784	12	-2.337	0.038*
S vs. No S	28	-0.036	0.972	26	3.175	0.004*
T vs. No T	21	-0.335	0.741	22	0.331	0.774

Accuracy. Fig.5 shows the mean and standard deviation of the number of errors, for the three pairs of multimodal combinations, i.e. with/without K, S and T. Results of t-tests shows that, for both navigation and editing tasks, a keyboard significantly increases and speech significantly decreases the accuracy of a user.

User Experience Level. Fig.6 shows the mean and standard deviation of the user experience levels graded by the participants, for the three pairs of multimodal combinations, i.e. with and without K, S and T. Table 7 shows the result of t-tests comparing the performance of the above mentioned multimodal combinations pairs for navigation and editing task. Results of the t-tests, showed that, for editing tasks, a keyboard significantly increases and touch significantly decreases the user experience.

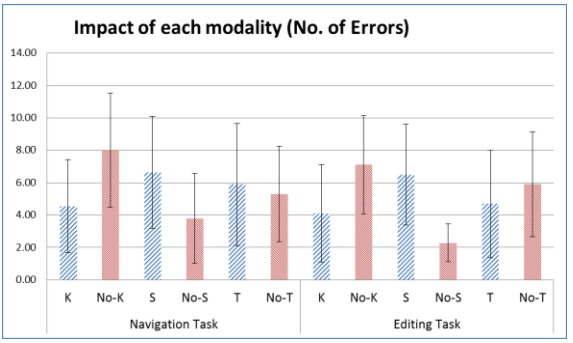

Fig. 5. Mean and S.D. w.r.t. number of errors with and without keyboard, speech and touch

Table 6. T-test results of number of errors with and without keyboard, speech and touch

	Navigation Task			Editing Task		
Multimodal pairs	df	t stat	sig. '*p*'(2-tailed)	df	t stat	sig. '*p*'(2-tailed)
K vs. No K	15	-2.676	0.017*	18	-2.560	0.020*
S vs. No S	22	2.435	0.023*	27	5.328	0.000*
T vs. No T	23	0.478	0.637	19	-0.946	0.356

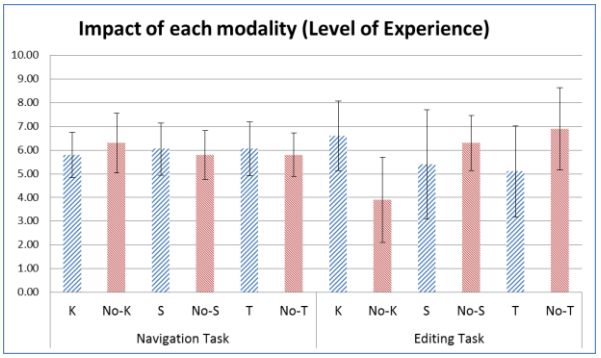

Fig. 6. Mean and S.D. w.r.t. experience levels with and without keyboard, speech and touch

Table 7. T-test results of user experience levels with and without keyboard, speech and touch

	Navigation Task			Editing Task		
Multimodal pairs	df	t stat	sig. '*p*'(2-tailed)	df	t stat	sig. '*p*'(2-tailed)
K vs. No K	14	-1.113	0.285	15	4.125	0.001*
S vs. No S	19	0.612	0.548	28	-1.424	0.166
T vs. No T	22	0.645	0.525	20	-2.591	0.017*

Our results showed that, the multimodal combinations with speech had low performance and accuracy. Few earlier works [5] [6] too reported similar results. Speech does contribute to the speed of task; however, due to less accuracy of speech this contribution is not achieved in a multimodal combination. Speech was found to be efficient for commanding and not for information entry. Keyboard was found to be efficient for information entry and not for commanding. Touch pen seemed good for pointing to entities which were easily located on the screen. In addition, it was found that the conventional interaction modality i.e. a keyboard has a significant role to play, even in the presence of non-conventional interaction modalities. Ren et al. [5] mentioned that mouse was useful and was more accurate than pen.

4.3 Are the usability parameters related?

Correlation was performed among the pairs of the three usability parameters used in the experiment, task completion times (representing performance), number of errors (accuracy) and user experience. Table 7 presents the Pearson's correlation coefficient '*r*' and corresponding significance value '*p*' between the pairs of usability parameters.

Table 8. Correlation among key usability parameters for multimodal combinations

(r, p)	Task completion times	Number of errors	User experience
Task completion times	1		
Number of errors	(0.626, 0.021)	1	
User experience	(-0.641, 0.009)	(-0.520, 0.041)	1

The results of correlation showed that, there is a strong correlation among the three usability parameters. This means, the multimodal combination having smaller task completion time (i.e. higher performance) is expected to have lesser number of errors (higher accuracy) and also higher user experience. Similar results were reported by Sauro and Kindlund, [15] who observed positive correlation between the direct data and indirect data from their experiments. Contrary to this, Hornbaek and Law [16] reported negative correlations between direct and indirect data.

5 Conclusions and Future Work

In this study, we have formulated few research questions on effectiveness of different multimodal input interactions, collected the user data through experiments, and performed analysis to answer those research questions. Speech input in general was observed to be the fastest input modality, but due to its low accuracy and uncertainty its performance gets compromised [17]. Speech input seems to be effective for commanding but not for information entry; unlike the keyboard which is more effective for information entry but not for commanding. Contrastingly, touch as well as mouse seems to be effective in pointing any visible GUI entity on the screen. Knowing such information related to the effectiveness of multimodal combination is crucial in

designing multimodal interactions. For any application which requires extensive use of GUI but minimal use of text editing, speech along with touch would be the preferable multimodal combination. Similarly, for an application requiring more text editing and lesser navigation, speech together with keyboard would be preferable.

The work presented in this paper is an initial research in the direction of developing quantitative models of multimodal combination, which could guide in designing the multimodal interactions for different applications. Future work would involve inclusion of other non-conventional input modalities like hand or body gesture in the experiment. This work may also be extended for specific domain applications, and more involved user groups like bank teller, tele-caller, industrial plant operator, etc.

References

1. Dybkaer, L., Bernsen, N.O., Minker, W.: New Challenges in Usability Evaluation-Beyond Task-Oriented Spoken Dialogue Systems. In: ICSLP, vol. 3, pp. 2261–2264 (2004)
2. Bretan, I., Karlgren, J.: Synergy Effects in Natural Language-Based Multimodal Interaction. SICS Research Report (1994)
3. Wechsung, I., Engelbrecht, K.P., Schaffer, S., Seebode, J., Metze, F., Möller, S.: Usability Evaluation of Multimodal Interfaces: Is the Whole the Sum of Its Parts? In: Jacko, J.A. (ed.) HCI International 2009, Part II. LNCS, vol. 5611, pp. 113–119. Springer, Heidelberg (2009)
4. Oviatt, S.: User-centered modeling and evaluation of multimodal interfaces. IEEE 91(9), 1457–1468 (2003)
5. Ren, X., Zhang, G., Dai, G.: An experimental study of input modes for multimodal human-computer interaction. In: Tan, T., Shi, Y., Gao, W. (eds.) ICMI 2000. LNCS, vol. 1948, pp. 49–56. Springer, Heidelberg (2000)
6. Metze, F., Wechsung, I., Schaffer, S., Seebode, J., Möller, S.: Reliable Evaluation of Multimodal Dialogue Systems. In: Jacko, J.A. (ed.) Human-Computer Interaction, Part II, HCII 2009. LNCS, vol. 5611, pp. 75–83. Springer, Heidelberg (2009)
7. Bernsen, N.O., Dybkjær, L.: Evaluation of spoken multimodal conversation. In: 6th ACM International Conference on Multimodal Interfaces, pp. 38–45 (2004)
8. Serrano, M., Nigay, L.: A wizard of oz component-based approach for rapidly prototyping and testing input multimodal interfaces. J. Multimodal User Interfaces 3(3), 215–225 (2010)
9. Bernhaupt, R., Palanque, P., Winckler, M., Navarre, D.: Usability Study of Multi-modal Interfaces Using Eye-Tracking. In: Baranauskas, C., Abascal, J., Barbosa, S.D.J. (eds.) INTERACT 2007. LNCS, vol. 4663, pp. 412–424. Springer, Heidelberg (2007)
10. Bernhaupt, R., Navarre, D., Palanque, P., Winckler, M.: Model-Based Evaluation: A New Way to Support Usability Evaluation of Multimodal Interactive Applications. In: Maturing Usability: Quality in Software, Interaction and Quality, pp. 96–122 (2007)
11. Cohen, P.R., Johnston, M., McGee, D., Oviatt, S., Pittman, J., Smith, I., Chen, L., Clow, J.: QuickSet: Multimodal Interaction for Distributed Applications. In: 5th ACM International Conference on Multimedia, pp. 31–40 (1997)
12. Kaster, T., Pfeiffer, M., Bauckhage, C.: Combining Speech and Haptics for Intuitive and Efficient Navigation through Image Databases. In: 5th ACM International Conference on Multimodal Interfaces, pp. 180–187 (2003)

13. Beelders, T.R., Blignaut, P.J.: The Usability of Speech and Eye Gaze as a Multimodal Interface for a Word Processor. In: Speech Technologies, pp. 386–404 (2011)
14. Microsoft Voice Recognition System, http://www.microsoft.com/enable/products/windowsvista/speech.aspx (last retrieved on February 26, 2013)
15. Sauro, J., Kindlund, E.A.: Method to standardize usability metrics into a single score. In: ACM SIGCHI Conference on Human Factors in Computing Systems, pp. 401–409 (2005)
16. Hornbæk, K., Law, E.L.C.: Meta-analysis of correlations among usability measures. In: ACM SIGCHI Conference on Human Factors in Computing Systems, pp. 617–626 (2007)
17. Cohen, P.R., Johnston, M., McGee, D., Oviatt, S.L., Clow, J., Smith, I.: The efficiency of multimodal interaction: A case study. In: International Conference on Spoken Language Processing, vol. 2, pp. 249–252 (1998)