ISCA Archive
http://www.isca-speech.org/archive

INTERSPEECH 2012
ISCA's 13th Annual Conference
Portland, OR, USA
September 9-13, 2012

Using Quality Ratings to Predict Modality Choice in Multimodal Systems

Ina Wechsung[1], *Klaus-Peter Engelbrecht*[1] *& Sebastian Möller*[1]

[1]QU Lab, Telekom Innovation Laboratories, TU Berlin, 10587 Berlin, Germany
ina.wechsung@telekom.de, klaus-peter.engelbrecht@telekom.de, sebastian.moeller@telekom.de

Abstract

A standardized procedure to evaluate the perceived quality of multimodal systems is still lacking. Previous research has however shown that the quality ratings for a multimodal system are equal to the weighted sum of the quality ratings of its individual modalities, with the modality that is more frequently used having a stronger influence. These findings suggest, that if the choice of modality can be predicted, an estimation of the quality of the multimodal systems is possible, based solely on an evaluation of its component modalities. Accordingly, the current study investigates the prediction of modality choice based on quality ratings of the component modalities, in order to achieve accurate quality predictions for multimodal systems. It is shown that predictions of modality choice as well as the overall system quality are possible. Furthermore, an age effect is observed: if older adults are included, predictions are less precise.
Index Terms: Multimodal systems, evaluation, prediction

1. Introduction and motivation

Systems offering multiple input and output interaction modalities have become increasingly widespread in recent years. But a standardized procedure to evaluate the perceived quality of such multimodal systems is still lacking. Previous research has, however, shown that the quality ratings for a multimodal system are equal to the weighted sum of the quality ratings of its individual modalities, with the modality that is more frequently used having a stronger influence [1], [2]. These findings suggest, that if the choice of modality can be predicted, then an estimation of the quality of the multimodal systems is possible based solely on an evaluation of its component modalities. Especially for early development stages such an approach could be beneficial in getting a rough approximation of the quality of a system without deploying complex evaluation procedures, which are often necessary for multimodal systems.

2. Related work - factors influencing modality choice

Several factors determining modality choice have been identified so far, with most studies focusing on efficiency [3],[4],[5],[6] and effectiveness [7],[8]. It has repeatedly been shown that a main factor influencing modality choice is not efficiency in terms of absolute task duration, but rather the relative efficiency. The relative efficiency is the amount of information that can be entered in one interaction step (bandwidth) in one modality, compared to the information bandwidth of an alternative modality [3],[4]. What has been reported in the literature is, the longer the input query, the more likely it is to be entered as speech input [3],[4]. These findings are in opposition to those of Kamvar and Beeverman, who showed that long queries are less likely to be entered with speech in a web search task [9].

But while the first two studies [3],[4] had a laboratory test set-up, the later one [9] used real-life log data and as such could not examine whether such long queries were entered using the copy and paste-function. If this were the case, then the GUI may well have been the most efficient modality, in terms of the previous definition.

For effectiveness, defined as the accuracy and completeness of goal achievement in the ISO standard [10], previous research did not observe a strong influence on modality choice: multiple studies report that at least for the initial correction attempt, users tended to stay in the erroneous modality and do not switch to another one [7],[8],[11],[12]. This might be due to the higher cognitive effort involved in modality switches, resulting in the development of a new problem-solving strategy compared to just reapplying same modality and problem solving strategy [12]. However, if users eventually do switch the modality, this is often due to the high level of error-proneness of the current modality, (i.e. insufficient effectiveness) [5].

Further to the classical usability concepts of efficiency and effectiveness, multiple additional factors have been identified. These are: the type of information [9], the prior knowledge about the system [13], the task [12],[14],[15] and the situational demands [16]. As one might expect, using speech input is less likely when entering confidential or private data [9]. Additionally, the stated modality preference is affected by the prior knowledge the user has about the system [13]. In a study by Jokinen and Hurtig [13], participants received either information that the tactile modality is supplemental to the speech modality, or that speech is supplementing the tactile modality. Both groups preferred the respective supplemental modality. Note that in this study it was the *stated* modality preference that was measured, rather than actual modality usage.

Regarding task characteristics, it has been shown that for navigational tasks the use of speech is less likely, when compared to touch or stylus input [12],[14],[15]. These task-specific modality preferences were reported for stationary contexts like desktop settings [12],[14], as well as for mobile contexts like PDA usage [15].

Situational demands are often related to the allocation of attentional resources. For example, in the automotive domain the visual channel is busy due to the task of driving which means that the amount of visual attention left for parallel tasks. (e.g. entering a destination into a navigation system), is very limited. According to multiple resources theory (MRT), a psychological model assuming multiple cognitive resources [17], it is likely that in such a situation the user will prefer speech input over other modalities, as speech requires less visual resources compared to a graphical user interface. The assumptions of MRT are supported by the findings of [14]. They report that speech was preferred over gestures in an in-car scenario, whereas

gestures were preferred while walking. 2D gestures, the input method requiring the most visual attention, were rarely used in either scenario. Furthermore, a study by Cox et al. [18] provides evidence for Wicken's model: to imitate a situation with the visual channel being busy, (e.g. walking or driving), they offered only limited visual feedback for a text creation task, and found speech being preferred over keyboard input. Additionally, Wechsung et al. [16] showed that while interacting with a mobile device, speech input increased where there was a visually demanding parallel task, while touch was used more often when an auditorily demanding concurrent task was presented. It is noteworthy, that although the proportion of speech usage increased with a visually distracting task compared to the proportion of the speech usage in auditory distracting task, touch was preferred over speech in both scenarios, regardless of the type of distraction introduced.

Based on the studies presented above, modality choice is influenced by a multitude of factors. According to the taxonomy of experience of multimodal human-machine interaction by Moeller et al. [19], all of the above factors also have an influence on the interaction with the system as well as the perceived quality of the system. Hence it might be possible to predict modality choice based solely on the interaction parameters and quality ratings of the constituent components.

To test these assumptions we included quality ratings and interaction data of three systems in multiple regression analyses, in order to predict modality choice.

3. Method

3.1. Systems and tasks

We included 3 different multimodal applications, each installed on a different smartphone, in our study. For all systems the available input modalities were speech and touch. Touch input always had to be entered with a swipe or tap gesture using the fingertips. Only the input modalities were varied; output was always multimodal and included feedback to the GUI, as well as task-specific auditory output. Apart from the "standard" output given for both modalities, special auditory feedback was provided for speech input for the following system states: recognition active, match, no match.

For the first study we used the *speech box* application, a multimodal mailbox system capable of handling speech-, email- and fax-messages, as well as of forwarding calls and mailbox message notification.

The application was installed on a HTC Touch Diamond. The speech module used was IBM Embedded ViaVoice. Speech recognition was activated via a push-to-talk button on the left hand side of the device. The experimental tasks were: accessing voice messages; retrieving a specific voice message; deleting this message; accessing the email inbox; opening an email; opening the fax inbox, opening a fax; opening the voice messages; sorting them from A to Z; redirecting all calls and confirming this change returning to the menu; and closing the *speech box*. For this system motion control was also implemented but was rarely used and thus excluded from the analysis.

The second study was conducted with a mobile *jukebox* application. It was installed on an HTC G1, an Android-based smartphone. The available input modalities were speech and touch. Speech recognition (Nuance Vocon) was activated via a push-to-activate button installed on the back of the device. Participants had to perform the following tasks: opening playlists, opening favorites, opening artist list, opening album list, searching for the song "first", for the song "second", for the song "third", for the song "fourth", and for the song "fifth", starting playback, skipping the current song, starting shuffle mode and stopping playback. Apart from the tasks explained above the participants had to respond to either visually or auditorily demanding stimuli which were presented in a randomized order in time intervals between 3 and 5 seconds.

System 3 was a multimodal *remote control* application for an IPTV service. It was implemented on an iPhone 3GS iPhone. The speech recognition Nuance Vocon was activated with a diagonal swiping gesture on the touch screen. Feedback was, depending on the task, either given via the TV or via the iPhone screen. The tasks included in the analysis were zapping, switching channels, increasing and decreasing the volume, setting and resetting time shift, opening teletext, retrieving a teletext site, opening and closing the EPG, starting and ending recording, retrieving information for the current program, activating and deactivating mute.

3.2. Participants

We included only complete cases in the further analysis. Accordingly the following descriptions do not include data of incomplete cases. In all studies none of the participants had any prior experience with the application and all participants were rewarded with either shopping vouchers or money in cash.

In the *speech box* study 23 German-speaking participants aged between 24 and 71 years (M = 43 y., SD =18 y.) took part.

In the *jukebox* study 24 German-speaking subjects, aged between 22 and 33 years (M = 26 y., SD =2 y.), participated.

For the multimodal *remote control* study we invited 32 German-speaking participants, aged between 18 and 62 years (M = 45 y., SD=12 y.). All were owners of the IPTV service tested.

3.3. Measures

We collected quality ratings with the AttrakDiff questionnaire [20]. In the first study we used the complete, 28 items version of the AttrakDiff. In the studies *jukebox* and *remote control* we employed the short version, containing 10 items. Both versions measure the perceived quality on 4 scales. These scales are *Hedonic Qualities - Stimulation (HQS), Hedonic Qualities - Identity (HQI), Pragmatic Qualities (PQ)* and *Attractiveness (ATT).* Hedonic qualities refer to the non-instrumental attributes of a system: The ability to evoke pleasure and emphasize the psychological well-being of the user. The scale *PQ* covers the classical usability attributes: the functionality and the access to the functionality. *ATT* is the global scale measuring both, hedonic and pragmatic qualities [20].

Moreover we assessed the perceived effort with the *SEA* scale [21]. The *SEA* scale (*Subjektiv Erlebte Anstrengung*) is a unipolar instrument ranging between 0 and 220 with higher values indicating higher effort.

Regarding interaction parameters we logged the task aborts after 3 unsuccessful attempts and task duration. To assess modality choice the modality chosen first to solve the task was logged for each task in the multimodal condition. Then the proportion of speech and touch was calculated. As participants could either choose touch or speech, the resulting usage rates were complementary, adding up to 100%. Consequently, for further analysis we used the proportions of speech only. Regression analyses using the touch usage rate are the same.

3.4. Procedure

For all studies participants had to sign a consent form and were asked to fill in demographic questionnaires. Next the applications were explained to them. In the unimodal condition all tasks had to be performed with either touch or speech. In the multimodal condition, participants could choose their preferred modality. The multimodal condition was always presented after the unimodal conditions. In order to avoid learning effects the sequence of the unimodal conditions (speech-touch vs. touch-speech) was altered for each participant in each study.

For each task the participants had 3 trials, after 3 trials the task was aborted and the next task had to be carried out. Interaction parameters were logged during the interaction. Quality ratings were assessed after each condition.

4. Results

In the presentation of results, sub-indices are used to indicate parameters collected in each condition, e.g. HQI_S for the rating on *HQI* scale in the *speech* condition, and HQI_T and HQI_{Mm} for the same ratings in the *touch* and *multimodal* conditions respectively.

4.1. System-wise prediction of modality choice

Stepwise linear regression with modality choice as dependent and the quality ratings and interaction parameters in the unimodal conditions as predictor variables was conducted for each system.

For the *jukebox*, HQI_S and SEA_S were included as predictors. If the ratings for speech on the *HQI* scale were high and the perceived effort of speech was low speech usage decreased. The inclusion of the *SEA* scale might be due to the concurrent task the users had to perform. For the *remote control* – similar to the *jukebox* – HQI_S was included in the model (Table 1).

For the system *speech box*, no significant predictor was found. In this experiment, the maximum age of the participants was higher compared to the other studies. Since previous research reported that prediction is difficult for older adults [1], possibly due to age-related decrease in memory capacity, we excluded participants older than 55 years.

With only the younger users, also for the *speech box* a significant predictor was found. But, in contrast to the other two systems, not HQI_S was included, but ratings on the global scale in the speech condition (ATT_S). Better ratings on ATT_S are related to increased usage of speech.

For the *jukebox* system, the exclusion of older users had no effect on the model, as all participants were younger than 56 years. For the *remote control*, the predictors remained the same (only HQI_S), but the accuracy increased considerably (Table 1).

4.2. Global analysis prediction of modality choice

In the next step we performed multiple linear regression analysis on the data of all 3 systems together. If older and younger users were included, HQI_S and HQS_T were significant predictors for modality choice. The better the ratings for *speech* on the *HQI* scale and the worse the ratings for *touch* on the *HQS* scale the more likely was the usage of speech in the multimodal condition. If older participants were excluded, HQI_S remained in the model, while HQS_T was removed. Instead, besides HQI_S, ratings on PQ_S and the abort rates in the *speech* condition ($Abort_S$) were chosen by the algorithm. Again prediction accuracy was higher without the older users (Table 2).

System	Older users	Predictor	β	F (df)	p	Adj. R^2	RMSE
Speech box	w	-	-	-	-	-	-
	w/o	ATT_S	.556	5.82 (1,13)	.031	.256	.24
Jukebox	w	SEA_S	-.625	8.17 (2,21)	.002	.384	.15
		HQI_S	.349				
Remote control	w	HQI_S	.466	8.34 (1,30)	.007	.191	.28
	w/o	HQI_S	.633	15.37 (1,23)	.001	.375	.21

Table 1. *Results of stepwise multiple linear regression analyses for each system.*

Older users	Predictor	β	F (df)	p	Adj. R^2	RMSE
w	HQI_S	.487	7.43 (2,76)	.001	.141	.28
	HQS_T	-.253				
w/o	PQ_S	.217	14.68 (3,60)	<.001	.395	.20
	HQI_S	.339				
	$Abort_S$	-.263				

Table 2. *Results of stepwise multiple linear regression analyses overall systems.*

4.3. Prediction of multimodal quality ratings based on modality choice predictions

As the prediction performance of all models above increased if older participants were excluded, we performed the following analyses with younger users only. To see if the predicted proportion of modality usage can be used to predict quality ratings in the multimodal condition (Q_{Mm}), the predicted values were used as coefficients to the ratings in the single modality conditions (Q_T and Q_S), as suggested by previous work outlined above. The resulting equation is as follows:

$$Pr\_Q_{Mm} = Pr\_Use_S * Q_S + Pr\_Use_T * Q_T \ , \qquad (1)$$

where Pr_Q_{Mm} is the predicted quality rating in the multimodal condition, Pr_Use_S is the predicted proportion of speech usage, and Pr_Use_T is the predicted proportion of touch usage. Note that Q may represent any of the 4 scales of the AttrakDiff.

The correlation between the quality ratings Q_{Mm} and the predictions with Equation 1, Pr_Q_{Mm}, was quite high for all scales (Table 3). To cross-check how much information was actually added by the predicted modality usage proportions, we further analyzed the correlation between Q_{Mm} and the mean of Q_S and Q_T, which corresponds to assuming equal distribution of modality usage in Equation 1:

$$Pr\_Q_{Mm} = 0.5 * Q_S + 0.5 * Q_T \qquad (2)$$

Table 3 shows that this model performs well, too. However, correlations were higher for all scales if the predicted modality proportions were included (Equation 1). A paired t-test confirmed that the absolute prediction error was significantly smaller for the scales *HQI* and *PQ* when using Equation 1. For the other scales, the error was smaller using Equation 1, but the difference was not significant.

Scale	Equation 1			Equation 2			t-test	
	Pr_Q_{Mm}	Abs. Error Pr_Q_{Mm}		Pr_Q_{Mm}	Abs. Error Pr_Q_{Mm}			
	Pearson's R with Q_{Mm}	M	SD	Pearson's R with Q_{Mm}	M	SD	p	t (63)
HQI	.828	.395	.41	.727	.455	.45	.016	2.19
HQS	.894	.402	.33	.859	.404	.36	.473	.069
ATT	.857	.431	.35	.834	.472	.45	.160	.968
PQ	.742	.544	.51	.716	.607	.55	.043	1.75

Table 3. *Correlation between predicted quality ratings and actual ratings and t-test for absolute errors*

5. Discussion and Conclusions

The study tried to predict modality choice based on quality ratings in order to achieve accurate quality predictions for multimodal system from data gathered in assessment studies for the component modalities. It was shown that such predictions of modality choice are possible and that with those predictions also the prediction accuracy of the overall quality of multimodal systems is improved. In practical terms, this means that an estimation of their perceived quality can be obtained from the ratings of their individual components without carrying out complicated multimodal evaluation experiments.

It was further observed that interaction parameter cannot explain modality choice; which is in line with the results of previous studies [23]. It has been reported that interaction parameters assessing efficiency and effectiveness do not necessarily reflect the perceived efficiency or effectiveness. Since judgments and decisions are made based on the individual perceptions and evaluations, it is reasonable to assume that the perceived efficiency and effectiveness are better predictors for modality choice. However, this is also not consistently the case for the current data. The scale *Pragmatic Qualities (PQ)* measuring quality attributes related to efficiency and effectiveness was less often included in the models than the scale *Hedonic Qualities - Identity (HQI)*. This means that modality choice is, aside from the factors explained in the introduction, not determined primarily by the modalities' instrumental qualities, but by its non-instrumental, hedonic qualities. Nevertheless, interaction parameters may have an indirect influence on modality choice moderated by the perceived quality ratings. Such indirect influences cannot be assessed with linear regression; hence these assumptions will be tested in a next step using structural equation modeling.

Also ratings for touch were seldom included in the models. A possible explanation might be that touch is the "default" modality, and only if speech is perceived as possessing high hedonic qualities the users actually use speech. However, all experiments included in the current study were lab experiments and it is reasonable to assume, that in the real world (e.g. in a work context), system pragmatic qualities can be more important. Accordingly these results have to be verified in more natural usage contexts. Furthermore. the systems investigated offered to a large extent sequential input only. If predictions, as described in the current paper, are possible also for systems offering extensive parallel input needs to be determined.

6. References

[1] Wechsung, I., Engelbrecht, K.-P., Schaffer, S., Seebode, J., Metze, F. and Möller S., "Usability Evaluation of Multimodal Interfaces: Is the Whole the Sum of Its Parts?", Proc. of HCII 2009 (2): 113-119, 2009.

[2] Wechsung, I., Engelbrecht, K.-P., Naumann, A., Schaffer, S., Seebode, J., Metze, F. and Möller S., "Predicting the quality of multimodal systems based on judgments of single modalities." Proc. of INTERSPEECH 2009: 1827-1830, 2009.

[3] Wechsung, I., Engelbrecht, K.-P., Naumann, A., Möller, S., Schaffer, S. and Schleicher, R. "Investigating modality selection strategies", Proc. SLT 2010, 31-36, 2010.

[4] Perakakis, M. and Potamianos, A., "A study in efficiency and modality usage in multimodal form filling systems." IEEE Trans. Audio Speech Lang. Process. 16(6):1194-1206, 2008.

[5] Bilici, V., Krahmer, E., Riele, S. and Veldhuis, R. "Preferred modalities in dialogue systems." Proc. ICSLP 2000, 727-730, 2000.

[6] Rudnicky, A.I., "Mode preference in a simple data-retrieval task." Proc. of HLT 1993, 364-369, 1993.

[7] Sturm, J. and Boves, L. "Effective error recovery strategies for multimodal form-filling applications." Speech Communication 45(3): 289-303, 2005.

[8] Suhm, B., Myers, B. and Waibel, A. "Multimodal error correction for speech user interfaces." ACM T COMPUT-HUM INT 8(1):60-98, 2001.

[9] Kamvar, M. and Beeferman, D. : "Say what? why users choose to speak their web queries", INTERSPEECH 2010: 1966-1969. 2010.

[10] International Standardization Organization (ISO) "9241-11, Ergonomic requirements for office work with visual display terminals (VDT) - Part 11 Guidance on usability", ISO, 1998.

[11] Lai, J., Mitchell, S., and Pavlovski C. "Examining modality usage in a conversational multimodal application for mobile e-mail access." Int'l Journal of Speech Technology 10(1): 17-30, 2007.

[12] Chen, X. and Tremaine, M. "Patterns of Multimodal Input Usage in Non-Visual Information Navigation." Proc. of HICSS 2006, 123.3, 2006.

[13] Jokinen, K. and Hurtig, T. (2006) "User expectations and real experience on a multimodal interactive system.", Proc. of Interspeech 2006, 2006.

[14] Althoff, F., McGlaun, G., Lang, M., and Rigoll, G., "Evaluating multimodal interaction patterns in various application scenarios", Proc. of GW 2003, Genova, 421-435, 2003.

[15] Gong, L. "Multimodal interactions on mobile devices and users' behavioral and attitudinal preferences." Proc. of HCII 2003, (4), 1402-1406, 2003.

[16] Wechsung, I., Schleicher, R. and Möller, S. "How Context Determines Perceived Quality and Modality Choice. Secondary Task Paradigm Applied to the Evaluation of Multimodal Interfaces." Proc. of IWSDS 2011, Springer, 327-342, 2011.

[17] Wickens, C.D., and McCarley, J. "Applied attention theory." Boca-Raton, FL, Taylor & Francis, 2008.

[18] Cox, A.L,, Cairns, P.A., Walton, A. and Lee, S. "Tlk or txt? Using voice input for SMS composition. Using voice input for SMS composition." Personal Ubiquitous Computing 12(8), 567-588, 2008.

[19] Möller, S., Engelbrecht, K.-P., Kühnel, C., Wechsung, I. and Weiss, B. "A Taxonomy of Quality of Service and Quality of Experience of Multimodal Human-Machine Interaction." Proc. of QoMEX 2009, 7-12, 2009.

[20] Hassenzahl, M., Burmester, M. and Koller, F. "AttrakDiff: Ein Fragebogen zur Messung wahrgenommener hedonischer und pragmatischer Qualität" [A questionnaire for measuring perceived hedonic and pragmatic quality], Proc. of Mensch & Computer 2003.Teubner, Stuttgart, Germany, 187–196, 2003.

[21] Eilers, K., Nachreiner, F. and Hänecke, K. "Entwicklung und Überprüfung einer Skala zur Erfassung subjektiv erlebter Antrengung." [Development and evaluation of a scale to assess subjectively perceived effort]. Zeitschrift für Arbeitswissenschaft, 40: 215-224, 1986.

[22] Engelbrecht, K.-P., Möller, S., Schleicher, R. and Wechsung, I. "Analysis of PARADISE Models for Individual Users of a Spoken Dialog System", Proc. of ESSV 2008, 86-93, 2008.

[23] Hornbæk, K. and Law, E.L. "Meta-analysis of correlations among usability measures." In Proc. CHI 2007, 617-626, 2007.