Chapter 9

QUALITY OF EXPERIENCING MULTI-MODAL INTERACTION

Benjamin Weiss, Sebastian Möller, Ina Wechsung and Christine Kühnel
Quality & Usability Lab, Deutsche Telekom Laboratories
TU Berlin, Germany
{BWeiss,Sebastian.Moeller,Ina.Wechsung,Christine.Kuehnel}@telekom.de

Abstract In this chapter, we discuss the contributions of different modalities to the overall quality of multi-modal interaction. After reviewing some common systematics and findings concerning multi-modality, we present experimental results from several multi-modal scenarios, involving different (human-to-human and human-to-machine) interaction paradigms, different degrees of interactivity, and different (speech, audio, video, touch, gesture) modalities.The results show that the impact of each modality on overall quality in interaction depends heavily on the scenario and degree of interactivity. Complementary modalities are not considered in this paper, but the models presented allow predicting overall system quality on the basis of individual modality ratings with an appropriate accuracy. These models still have to be validated in order to be used as tools for system developers estimating whether adding modalities will have an impact on the quality experienced by the user.

Keywords: Usability; User experience; Perceived quality; Multi-modal integration.

1. Introduction

Multi-modal dialog systems appear to offer better interaction experience, as multi-modality seems to have fundamental advantages over unimodal interaction. However, there are few matching examples beyond the standard "put-that-there" scenario. Much more often, simply providing alternative input or output modalities resulting in sequential multi-modality seems to be the state-of-the-art. The question is what constitutes a "good" interaction, i.e. what aspects contribute to the user having a good or bad impression of the system she has been using. This is commonly understood by the term "Quality of Experience", QoE.

W. Minker et al. (eds.), *Spoken Dialogue Systems Technology and Design*,
DOI 10.1007/978-1-4419-7934-6_9,

In this chapter, we will summarize major results concerning multi-modality at first, and then provide a common ground on what Quality of Experience really means. We will then present experimental results from different interaction scenarios: Audio-visual transmission systems (like IP-based television or audio-visual telephony), interactions with Embodied Conversational Agents (ECAs), as well as interactions with different non-embodied multi-modal dialogue systems providing speech, touch and motion input capabilities. For each scenario, algorithmic models are presented which quantify the impact of each modality on the overall system quality, as it is perceived subjectively by the user. The goodness of the models are described in term of Pearson's correlation R between the models' estimates and the real data obtained, as well as the root mean squared prediction error (RMSE). We conclude by identifying some research questions which should be answered in order to fully support the design and evaluation of multi-modal dialog applications.

2. Advantages of Systems Providing Multi-Modal Interaction

One major assumption concerning human-computer interfaces is that the interaction is significantly facilitated by providing multiple input modalities and by presenting information over different output channels. From a usability point of view – i.e. discounting hedonic aspects like appearance and style of the interface or the possibility to express the user's identity with a given product – a multiple of possible input modalities can increase the recognition rate by fusing different input modalities (e.g. on the signal level) and it allows people to use those modalities most adequate in their specific situation, mood and capability (López-Cózar Delgado and Araki, 2005; Oviatt, 2004). For example, touch may be favoured in noisy or public environments, speech for the task of selecting objects in longer lists, typing for editing text and pointing gestures to refer to spatial information. Concerning the system output, multiple modalities allow for selecting the most appropriate way to present a specific piece of information (e.g. Graphical User Interfaces for lists, Embodied Conversational Agents for emotions, auditory icons for alarms, short vibrations for positive feedback). Another benefit is the possibility to present information redundantly to increase salience.

Furthermore, there seem to be cognitive advantages for multi-modal interfaces. Redundant and complementary information may distribute the use of cognitive resources and thus make processing faster and less demanding. With the theory of multiple resources (Wickens, 1999) for example, the tasks of speaking and gesturing or hearing and watching use different resources that in principle should not interfere with each other. As a result, users seem to prefer multi-modal interaction, especially, when the cognitive load increases

due to time pressure or task difficulty (cf. (Oviatt et al., 2004)). However, there are also examples which show that this benefit is not always observable, and the theoretical basis of a strict separation of the unimodal signals is questioned (cf. (Sarter, 1995)). Instead, multi-modality may even increase cognitive load (Schomaker et al., 1995) compared to single-modality usage.

As humans naturally interact with each other multi-modally – i.e. face-to-face communication with speech, non-speech sounds, gestures, expressions – an Embodied Conversational Agent used adequately as an interface to computers can increase user's experience of a system (cf. (Benoît et al., 2000)). Of course, this also holds for communication services enabling human-to-human interaction with more than one modality, commonly by providing audio-visual (AV) communication.

Certainly, multi-modal interfaces enable a new quality of human-to-human and human-to-machine interaction. To achieve this expected benefit in user experience, we have to know how users experience the interaction with such systems and services. In the following, relevant mechanisms are explained showing how users come to their judgments of system quality and how different modalities contribute to this. For three different scenarios, namely multi-modal signal transmission, Embodied Conversational Agents (ECAs), and non-embodied dialog systems, experimental results are summarized to derive simple algorithmic models of the integration processes for overall quality ratings of multi-modal systems. It will be pointed out which problems have to be dealt with and what steps have to be taken in order to really predict the quality users experience when interacting with multi-modal dialog systems.

2.1 Modality Relations

There are different approaches to formalize the relationship between different modalities during an interaction. Typically, there are two dimensions addressed: The *temporal assignment* (parallel vs. sequential multi-modality) and the *amount of information conveyed* with each modality (complementary vs. redundant). One of the most common systematics is described by the CARE properties (Coutaz et al., 1995). Apart from formal definitions of the name-giving four properties, the relationship between the multi-modal behavior of the user and the one of the system is discussed:

- **C**omplementarity: Different modalities have to be used in order to reach the target.
- **A**ssignment: Only one modality is selected, either by the system or the user.
- **R**edundancy: Different modalities are used, bearing comparable information, either in parallel or sequentially.

- **E**quivalence: Any available modality can be used. There are no restrictions on the temporal order.

Such formal descriptions of modality relations can be used to specify how a multi-modal system outputs information generically, or dependent on the user input. For specified tasks and user groups this formalization can also be used for evaluating the system's appropriateness in modality choices. But also face-to-face and thus human-to-ECA communication might be formalized by this account. It is not trivial to simulate human behavior with ECAs, as linguistic and non-linguistic information might be naturally redundant (e.g. mood is expressed with voice as well as facial expressions and posture), but information often conveyed complementarily (the famous "put-that-there" scenario).

3. Quality of Experience

Developers of multi-modal systems tend to highlight the performance of their system and the individual input and output modules in order to justify how good their system is. In this context, we can define "performance" as follows:

Performance: *The ability of a unit to provide the function it has been designed for (Möller, 2005).*

Easy-to-calculate performance figures are e.g. the recognition rates for speech or gesture recognizers, the intelligibility of TTS modules, or the conveyability of intended emotions by an ECA. A pre-defined set of performance figures can be used to characterize the so-called "Quality of Service", QoS. This term which is commonly used for media transmission services has been defined as follows:

Quality of Service (QoS): *The collective effect of service performance which determines the degree of satisfaction of the user of the service (ITU-T Rec. E.800, 1994). This includes service support, service operability, serveability, and service security.*

Although system performance (and thus QoS) will have a severe impact on user satisfaction, there is no one-to-one relationship between the two. User satisfaction is just one aspect of quality, i.e.:

Quality: *Result of appraisal of the perceived composition of the service with respect to its desired composition ((ITU-T Rec. P.851, 2003), following (Jekosch, 2004; Jekosch, 2005)).*

Apparently, quality requires a perception and a judgment process to take place inside the human user. Obviously, the result of this process is severely impacted by the system characteristics (and so system performance), but there are

other characteristics of the usage situation and context as well as user-internal factors (memory, expectation, etc.) which will decide on which level of quality the user finally attributes to the interaction with the system. As a corresponding concept to Quality of Service, the term "Quality of Experience" (QoE) is now in use to summarize the user perceptions resulting from the interaction with the system. Unfortunately, QoE is still ill-defined in the international bodies:

Quality of Experience (QoE): *The overall acceptability of an application or service, as perceived subjectively by the end user. Quality of Experience includes the complete end-to-end system effects (client, terminal, network, services infrastructure, etc.) (ITU-T Rec. P.10, 2007).*

However, overall acceptability may be influenced by user expectations and context. A better definition emerged from discussions by the participants of the Dagstuhl Seminar 09192 "From Quality of Service to Quality of Experience" which was held in May 2009 in Dagstuhl, Germany:

Quality of Experience (QoE): *Degree of delight of the user of a service. In the context of communication services, it is influenced by content, network, device, application, user expectations and goals, and context of use.*

Service: *An event in which an entity takes the responsibility that something desirable happens on the behalf of another entity.*

Acceptability: *Characteristic of a service describing how readily a person will use the service. Acceptability is the outcome of a decision which is partially based on the Quality of Experience.*

In order to assess Quality of Experience, perception and judgment processes have to take place inside a human user. As a consequence, subjective evaluation methods are necessary in order to quantify the QoE which can be achieved with a particular multi-modal system. In (Möller et al., 2009), we have shown that QoE is a multidimensional construct, the components of which can be quantified with the help of dedicated questionnaires. Table 1 is taken from (Möller et al., 2010) and summarizes some commonly used questionnaires which have been proved adequate to quantify sub-aspects of QoE, and which will be used in some of the studies cited and summarized hereafter.

4. Audio-Video Quality Integration in AV-Transmission Services

In the case of network services providing audio-visual signals like television, video-clips and especially AV-telephony, the perceived quality of the signals is one of the main factors to be assessed. Evaluating the visual and audio

Table 1. Comparison of questionnaires and captured QoE aspects[a]. ●: completely captured; ◐: partially captured; ○: not captured.

Sub-scales	*Questionnaire*			
	SUS	AttrakDiff[1]	SUMI[2]	SASSI[3]
Learnability	●	◐(PQ)	●(LEA)	●(LIK, HAB)
Effectiveness	●	●(PQ)	◐(CON, HEL)	◐(ACC, HAB)
Efficiency	●	●(PQ)	●(EFF)	◐(SPE, CD)
Intuitivity	○	○	○	○
Aesthetics	○	●(HQ-S, ATT)	◐(AFF)	○
System Personality	○	◐(HQ-S)	○	◐(ANN, LIK)
Appeal	○	●(HQ-S, HQ-I)	◐(AFF, LIK)	◐(ANN, LIK)

[a] Cf. (Möller et al., 2010) ©Elsevier 2010.

channel independently does not necessarily provide an insight into the quality experienced by the user. Instead, the mechanism of integrating both modalities during perception and appraisal has to be known in order to monitor and adjust the service.

4.1 Videotelephony

With a straight-forward approach, the perceived multi-modal quality (MOS_{AV}) is evaluated and modelled as a combination of the separate uni-modal quality ratings (MOS_A and MOS_V). Here, each rating is obtained on a 11-point Absolute Category Rating (ACR) scale as it is specified in (ITU-T Rec. P.920, 2000); then, Mean Option Scores (MOS) are derived for the audio, video and audio-visual quality of the transmission, by averaging the individual ratings over all users of the different test conditions. In this experiment, 24 subjects (aged 18–30 years) had to do the building block task (ITU-T Rec. P.920, 2000) and the short conversation test (Möller, 2000) via AV dialog. Three different simple relationships were tested:

$$MOS_{AV} = c_1 \cdot MOS_A + c_2 \cdot MOS_V + c_3, \tag{9.1}$$

$$MOS_{AV} = c_1 \cdot MOS_A \cdot MOS_V + c_2, \tag{9.2}$$

$$MOS_{AV} = c_1 \cdot MOS_A + c_2 \cdot MOS_V + c_3 \cdot MOS_A \cdot MOS_V + c_4. \tag{9.3}$$

With the second model (9.2) correlations between estimated and measured MOS_{AV} between $R = 0.93$ and $R = 0.99$ could be obtained.[4] As shown in Table 2, models with an interaction term describe the perceptual integration better than simple linear models.

As expected, the visual channel contributes stronger to the multi-modal quality ratings than the auditory channel. Therefore, the correlation between

Table 2. Modelling audio-visual integration for videotelephony[a].

Model	$MOS_{AV} =$	*Pearson's R*	*RMSE*
Linear:	$0.677 + 0.217 \cdot MOS_A + 0.888 \cdot MOS_V$	0.96	0.53
Interaction:	$1.3 + 1.1 \cdot MOS_A \cdot MOS_V$	0.99	0.95
Complete:	$0.517 + 0.0058 \cdot MOS_A + 0.654 \cdot MOS_V + 0.042 \cdot MOS_A \cdot MOS_V$	0.97	0.57

[a] Cf. (Belmudez et al., 2009) ©IEEE 2009.

MOS_V and MOS_{AV} is higher than between MOS_A and MOS_{AV} (see Figure 1). Within these models, the variance of MOS_{AV} is basically determined by the variance of MOS_V alone. Notably, the impact of audio quality increases with that of the video. Apparently, MOS_V comes first, but with better video quality, there is a perceptual saturation effect, and audio quality gets more important for the test participants. However, the exact weighting depends on the type of task and the degree of interactivity (passive test from literate vs. dialog). In the short conversation test, the audio plays a crucial role to fulfill the task: Quality in conditions with bad audio quality is rated significantly worse than in the building block task with comparable conditions. These are two very important context effects, that are of strong influence in all multi-modal interaction scenarios, as shown in the next sections.

With all these models presented, there is always the problem of collecting valid data from the test participants: Ideally, unimodal ratings should be assessed separately from the multi-modal condition. However, in some of the scenarios presented, this is not practical or even impossible (e.g. rating visual quality for the short conversation test or articulating ECAs). Most importantly, rating scales are not used in a linear way: For example, there are saturation effects of the scale itself, and categories of the ACR-scale used in the experiments presented do not match the idea of continuous quality ascription. With a more linear scale the models might benefit from a better description of the ratings obtained.

4.2 IP-Television

In the case of quality of IP-based TV quality assessment, results were first transformed to a so-called "perceptual scale" (R-scale, cf. (ITU-T Rec. G.107, 2005)) which is used as a basis for transmission planning models by the International Telecommunication Union, ITU-T. This scale is thought to avoid some of the non-linearities of the ACR scales used in the experiment. In a study by Garcia and Raake (2009), two different modelling approaches were evaluated. Both models estimate AV-quality (Q_{AV}) on the R-scale: On the one hand using single modality quality ratings as in the data of Belmudez et al. (2009) (quality

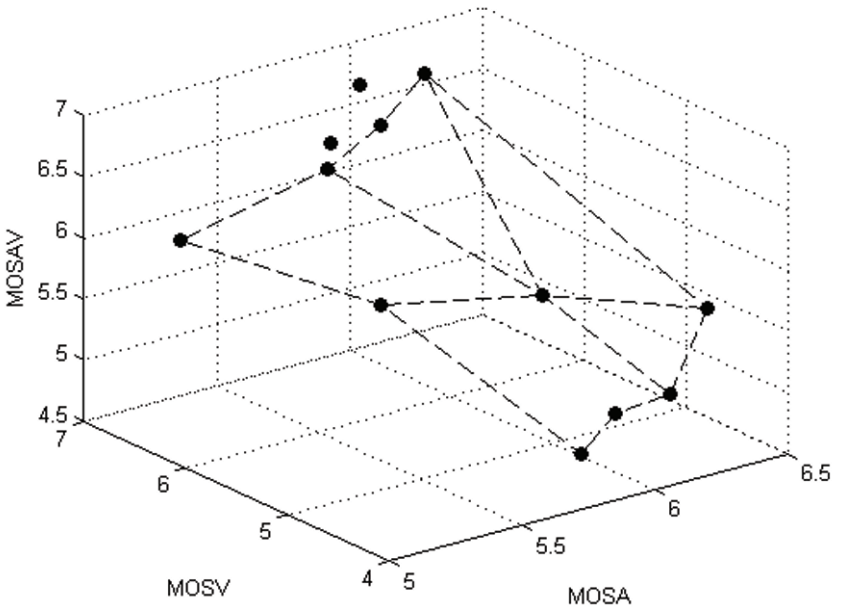

Figure 1. AV quality integration in video transmission services (from (Belmudez et al., 2009)), ©IEEE 2009.

based approach with the complete model, see Equation 9.4), and on the other hand an estimation of audio-visual quality based on impairments factors (impairment factor based approach, see Equation 9.5). For their data, impairment factor have been estimated from the subjective ratings, not from parametric descriptions of the transmission (e.g. packet loss). MOS are obtained from 24 different subjects (aged 21–44) for each of the three conditions: Audio-only, video-only and audio-visual. Both approaches show comparable results (see Figure 2).

$$Q_{AV} = 27.805 + 0 \cdot Q_A + 0.129 \cdot Q_V + 0.006 \cdot Q_A \cdot Q_V \qquad (9.4)$$

The quality-based model correlates with the subjects' ratings with $R = 0.96$ ($RMSE = 3.38$):

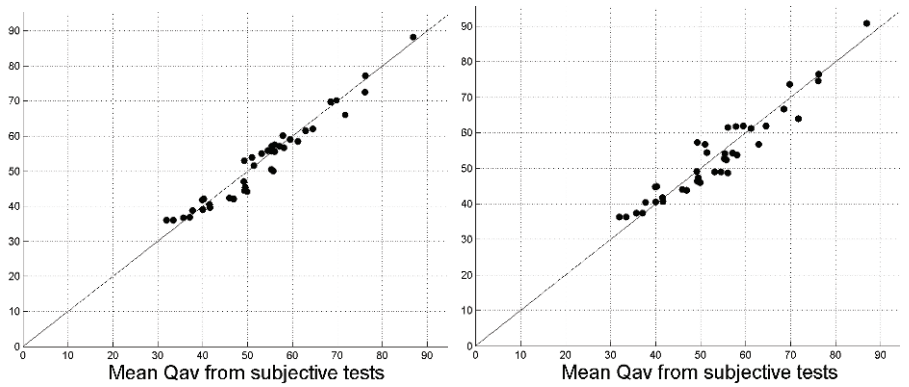

Figure 2. AV quality integration in IPTV. left: Quality-based approach, right: Impairment factor based approach (from (Garcia and Raake, 2009)), ©IEEE 2009.

$$\begin{aligned} Q_{AV} = 88.195 &- 0.379 \cdot Icod_A - 0.588 \cdot Icod_V \\ &- 0.625 \cdot Itra_A - 0.625 \cdot Itra_V \\ &+ 0.005 \cdot Icod_A \cdot Icod_V \\ &+ 0.007 \cdot Itra_A \cdot Itra_V \\ &+ 0.011 \cdot Icod_V \cdot Itra_A \\ &+ 0.007 \cdot Icod_A \cdot Itra_V. \end{aligned} \tag{9.5}$$

The impairment-based model performs slightly better ($R = 0.98$, $RMSE = 2.57$).

5. Quality of Embodied Conversational Agents

Dialogue systems with Embodied Conversational Agents are frequently represented by 3D modelled animated humans heads. Other realizations span from abstract icons (e.g. "smiley" faces) to animals, cartoons or fictional creatures. Concerning realistic human appearances, there are also visual models of the full body and upper part of the body in use. Apart from application in virtual realities, such an ECA can offer a number of benefits to a dialog system, including:

- intuitively display emotions and feedback (e.g. system state is idle, concentration on one of several user, system is busy);
- display of facial expressions or gestures for paralinguistic and linguistic usage;
- supporting the user to concentrate on the human-computer interface;

- increased robustness in speech perception (with lip-synchronous ECAs, e.g. in noise);
- general *Persona Effect* of better subjective ratings (cf. (Dehn and Van Mulken, 2000) for a meta-analysis and summary).

In a series of experiments, audio-visual quality of different talking heads was evaluated and modelled from single modality ratings of speech quality and visual quality (Weiss et al., 2010). In this case, different text-to-speech and head modules were used (see Figure 3 for pictures of the three talking heads tested). Transmission quality is not in scope of this research. Therefore AV quality of the heads presented on a display were comparable concerning codec and frame-rate. Instead, the perceived user experience of the talking head component was assessed as basis for the usability in their specific application. Subjects in the experiments rated the ECAs on several scales to assess various quality aspects.

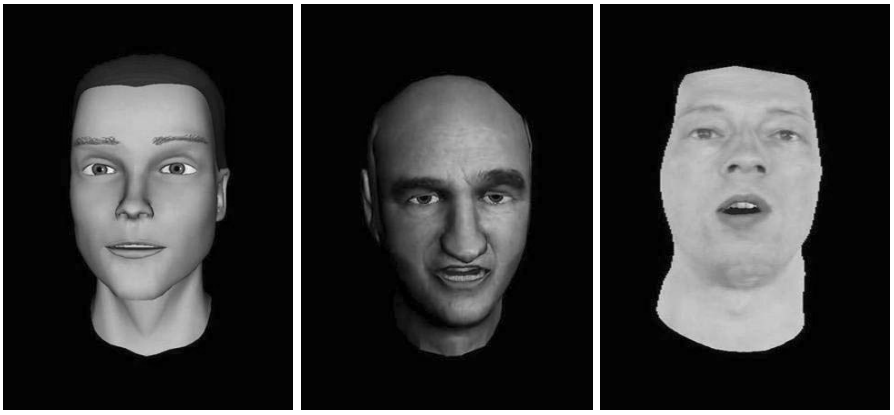

Figure 3. Three facial models tested (cf. (Kühnel et al., 2008)).

As a result of the experiments, talking heads overall quality (MOS_{heads}) could be described as a linear combination of visual MOS_V and speech quality rating MOS_A. However, the models do not perform comparable to AV quality assessed in IPTV or IP-videotelephony scenarios. The data sets obtained and described by the models are not truly comparable due to differences in stimuli. However, the most important results can be extracted from the models presented in Table 3. There were four different conditions: the passive rating test (14 subjects, aged 20–32), a simulated interaction (23 subjects, aged 21–60), a simulated interaction with an information screen in addition to the ECA screen (23 subjects, aged 20–57), and a real interaction with the system (49 subjects,

aged 20–61): With an increase in interactivity and an increase in distraction, module differences get blurred and the models' fit decreases. The distraction of the rating process was introduced by a second screen (the first displays the ECA). This additional screen presented information from the system as lists, whereas the degree of interactivity merely refers to the difference between a passive rating test versus a simulated interaction experiment. The last model is from a real interaction experiment which also included the second screen (cf. (Kühnel et al., 2009)).

Table 3. Modelling audio-visual quality of talking heads for passive, simulated (one and two screens), and real interaction scenarios (2 screens)[a].

type of experiment	$overallquality_{heads} =$	*Pearson's R*	*RMSE*
Passive:	$0.47 + 0.51 \cdot MOS_A + 0.33 \cdot MOS_V$	0.83	0.49
Simulated (1 screen):	$0.16 + 0.42 \cdot MOS_A + 0.30 \cdot MOS_V$	0.71	0.59
Simulated (2 screens):	$0.35 + 0.36 \cdot MOS_A + 0.23 \cdot MOS_V$	0.57	0.65
Interaction:	$0.30 + 0.26 \cdot MOS_A + 0.40 \cdot MOS_V$	0.57	0.62

[a] Cf. (Weiss et al., 2010) and (Kühnel et al., 2009).

From the questionnaires used it became clear, that at least for the data obtained during interaction, ECA quality cannot be sufficiently equated with auditory and visual quality. Other factors have also an significant impact on overall quality of talking heads: I.e. overall system quality, how entertaining the embodiment is due to non-linguistic movements, naturalness of the ECA, as well as perceived goodness of synchronization (cf. (Weiss et al., 2009)).

6. Quality of Systems with Multiple Input Modalities

The last scenario presented here are multi-modal dialog systems which can be controlled by different input modalities. Like current commercial systems, the experimental setup does allow to change the input modality sequentially, the appropriate CARE property is *Equivalence* (see Chapter 2.1). The questions addressed here are two-fold:

1 *Which modality is preferred by the user? How consistent is the individual modality usage?*

2 *Is a multi-modal interface better than an unimodal one? For systems providing* Equivalence*, can the multi-modal systems quality be modelled by ratings of the unimodal interfaces?*

Two experiments studying these issues are presented in the following: One interface is attached in an office area and one is a mobile device.

6.1 Smart Office

The so called *Attentive Display* is a room information system installed at T-Labs, Berlin. It provides information on the colleagues currently present and their desk and room bookings. Additionally, you can be informed about events (lectures, meetings). The interface is a big screen, fixed in the entrance area. It can be operated by touch and/or speech (that is enabled automatically when the camera tracks a face, thus the name "attentive display"). The output is always visual.

In the first experiment, there were three blocks: Touch only, speech only, and the multi-modal session always at the end of the test (cf. (Wechsung et al., 2009b) for the full description and results). User experience ratings were assessed with the AttrakDiff questionnaire (Hassenzahl et al., 2003), that covers hedonic and pragmatic aspects of the users perception (36 subjects, aged 21–39). For the three conditions, differences in the ratings on the hedonic and pragmatic scales were observed: The ranking concerning the pragmatic quality was touch over multi-modal over speech. This means, there was no benefit of providing speech in addition to touch. As subjects were explicitly asked to use the system multi-modally in the last session, the speech usage lead to lower ratings. However, the results are different for the hedonic scales: Here, multi-modal interaction was rated best. Please note, that the *Pragmatic* scale can be interpret as indicator of functionality and usability, whereas the global scale *Attractiveness* is related to user experience! Additionally, there is overall quality, which is the mean of all items used.

Concerning the integration of the quality ascribed with different input modalities, the multi-modal quality can be described as linear combination of the single modality ratings (see Table 4). As you can see, the fit of the models is far better for the overall quality and the *Attractiveness* scale. Mostly, touch is more important than speech – especially for *Attractiveness*, except for the *Identity* scale.

Table 4. Integration of perceived quality aspects of speech (Q_S) and touch (Q_T) to multi-modal ratings (Q_{MM}, ordered last)[a].

Scale	$Q_{MM} =$	*Pearson's R*	*RMSE*
Overall:	$0.14 + 0.81 \cdot Q_T + 0.68 \cdot Q_S$	0.91	0.35
Attractiveness:	$-0.20 + 0.85 \cdot Q_T + 0.48 \cdot Q_S$	0.92	0.41
Pragmatic:	$0.22 + 0.80 \cdot Q_T + 0.47 \cdot Q_S$	0.79	0.67
Stimulation:	$0.11 + 0.69 \cdot Q_T + 0.63 \cdot Q_S$	0.83	0.51
Identity:	$0.38 + 0.28 \cdot Q_T + 0.66 \cdot Q_S$	0.78	0.53

[a] Cf. (Wechsung et al., 2009b).

The multi-modal condition was always presented last in order to have the subjects become familiar with both modalities before using them together. A possible explanation for the observed strong correlation between the linear combination of both single modality ratings and the multi-modal condition could be that the subjects tried to rate consistently. To verify the results obtained, a second study was conducted with the multi-modal condition always at the first position (cf. (Wechsung et al., 2009a)). For this experiment the models extracted are significantly lower in power and stability (leave on out cross-validation, 18 subjects, aged 22–30). See Table 5 for the results. The *Pragmatic* and *Stimulation* scales are not included, as an estimation on basis of the single modality ratings was not possible.

Table 5. Integration of perceived quality aspects of speech (Q_S) and touch (Q_T) to multi-modal ratings (Q_{MM}, ordered first)[a].

Scale	$Q_{MM} =$	*Pearson's R*	*RMSE*
Overall:	$0.18 + 0.679 \cdot Q_T + 0.553 \cdot Q_S$	0.76	0.55
Attractiveness:	$0.29 + 0.653 \cdot Q_T + 0.545 \cdot Q_S$	0.77	0.73
Identity:	$0.06 + 0.664 \cdot Q_T + 0.485 \cdot Q_S$	0.87	0.41

[a] Cf. (Wechsung et al., 2009a).

6.2 Mobile

The application tested is a multi-modal information-box (e-mail, SMS, fax), that runs on a smart-phone (cf. (Wechsung et al., 2009a)). In addition to touch and speech, there is a motion input modality to navigate and select with tilting the whole device. The system's output is generally *assigned* to visual output. Additionally, there is context dependent *Redundancy* (cf. Section 2.1 for the CARE properties) for speech input (audio and visual): This is vibration as positive feedback for the motion modality and audio feedback to signal *match* and *nomatch* for voice input. The procedure is similar to the first experiment presented in the last section (30 subjects, two age groups: 25–29, 55–66): The multi-modal condition is always last. Findings include the relevance of the frequency of each modality used in the multi-modal condition: Motion is not included in the regression models (7% usage), and speech (19%) only for *Stimulation* (see Table 6). Combinations of modalites were only used infrequently (6%).

With a leave one out cross-validation the scale *Pragmatic* was identified as being unstable.

Table 6. Integration of perceived quality aspects of speech (Q_S), touch (Q_T) and motion (not significant) to multi-modal ratings (Q_{MM}, ordered last)[a].

Scale	$Q_{MM} =$	*Pearson's R*	*RMSE*
Overall:	$0.16 + 0.69 \cdot Q_T$	0.69	0.48
Attractiveness:	$0.04 + 0.79 \cdot Q_T$	0.56	0.68
Stimulation:	$0.31 + 0.60 \cdot Q_T + 0.35 \cdot Q_S$	0.86	0.40
Identity:	$0.22 + 0.75 \cdot Q_T$	0.69	0.45
Pragmatic:	$0.41 + 0.49 \cdot Q_T$	0.36	0.77

[a] Cf. (Wechsung et al., 2009a).

6.3 Summary

With the three experiments presented here, it could be shown that perceived quality aspects – including pragmatic and hedonic aspects – of multi-modal interaction could be described as linear combination of ratings for single modalities. Results are satisfying for overall quality and *Attractiveness*. Apparently participants are better in mentally "adding" than in "subtracting" modality ratings during evaluation, as subtracting one's own ratings from memory is more demanding (Kamii et al., 2001). This interpretation is supported by the finding, that older users – who often have decreased working memory capacity – multi-modal ratings are less good predicted. Interestingly, this became especially obvious for the *Pragmatic* scale.

While multi-modal conditions did not perform better than the best unimodal condition for the *Pragmatic* scale, on hedonic scales the quality did benefit from multi-modality. The amount of modality usage affects the weights of the single modalities.

7. Conclusions

Multi-modal communication systems can be found in a great variety of application scenarios. We presented evaluation experiments from fields of IP based audio-video transmission for TV and videotelephony, Embodied Conversational Agents for smart-home environments and stationary and mobile non-embodied multi-modal user interfaces. We showed how to assess perceived quality and user experience of such systems. Our results show that quality of multi-modal systems comprises a multitude of aspects – depending on the application – and is influenced by the measurement process.

For the case of audio-visual integration in video transmission applications, visual quality mostly has a much stronger influence on overall quality than audio/speech quality. Stable models with sufficient power can be derived for AV quality on the basis of single modalities' quality. However, the exact weightings depend on interactivity.

For audio-visual integration in ECA applications, interactivity also plays an important role: The degree of interactivity determines the impact of animation and speech on overall quality of the animated agent, but definitely other factors affect the ECAs overall quality as well, like the smoothness of interaction and other representations of the system. For example, additional information nicely presented by the system improved the ratings of the ECA.

Multi-modal quality and attractiveness of multi-modal interactive systems can be estimated on the basis of judgments for unimodal conditions. Complementary multi-modality ("put-that-there" scenario) was not tested, but are considered not common in commercial interactive systems. Weightings for overall quality reflect modality usage to a certain extent. Interestingly, weightings for hedonic qualities are also influenced by less-used modalities.

In all cases, the quality user are experiencing was assessed by questionnaires. To find models predicting perceived quality is difficult indeed: What kind of constructs (quality aspects) are relevant in the specific case and how are they assessed best? The AttrakDiff has shown great potential to cover many important aspects in a valid and reliable way for interactive systems. In the case of IP based transmission applications a continuously scale is recommended. But currently only some important factors have been identified to be included into the models or at least to be controlled in the experiments. The order of presentation of modalities and degree of interactivity are stated in this text, but of course the progress of interaction and topic of transmitted signals are relevant, too. There is a bunch of open questions, regarding this topic:

- If modality weightings are influenced by modality usage, what does influence actual modality usage?
- What is the impact of modality effectiveness and efficiency?
- For interactive systems, what is the impact of output modalities for the usage of input modalities and the multi-modal quality judgment?
- What type of model (linear, multiplicative, other nonlinear) is most adequate for multi-modal quality prediction?
- For which (input and output) modalities does such modeling work well?

In all scenarios presented, weighted combinations of ratings for single modalities (either in unimodal conditions or as separate ratings for multi-modal conditions) could be used to describe multi-modal quality of experience. Apparently, estimating multi-modal quality works best for transmission quality and can be used for prediction already. For this case, it seems, there are not as many possibly influencing factors as in dialog system, as those include the interactive part when rated. In Human Computer Interaction scenarios, findings are obtained revealing fundamental mechanisms influencing the judgment

process for multi-modal interaction. In both cases, however, we have far to go to model any real cognitive processes.

Acknowledgments

We would like to thank Benjamin Belmudez, Marie-Neige Garcia and Alexander Raake for providing their data and figures on IPTV and videotelephony.

This work was partly supported by the Deutsche Forschungsgemeinschaft DFG (German Research Community), grant MO 1038/6-1.

Notes

1. Sub-scales: *Attractiveness* (ATT), *Hedonic Qualities – Identity* (HQ-I), *Hedonic Qualities – Stimulation* (HQ-S), *Pragmatic Qualities* (PQ).
2. Sub-scales: *Affect* (AFF), *Control* (CON), *Efficiency* (EFF), *Learnability* (LEA), *Helpfulness* (HEL). SUMI is generally not recommended for evaluating multi-modal systems.
3. Sub-scales: *System Response Accuracy* (ACC), *Annoyance* (ANN), *Cognitive Demand* (CD), *Habitability* (HAB), *Likeability* (LIK), *Speed* (SPE).
4. With constant factors of $c_1 = 0.107 \ldots 0.121$ and $c_2 = 1.1 \ldots 1.5$

References

Belmudez, B., Möller, S., Lewcio, B., Raake, A., and Mehmood, A. (2009). Audio and Video Channel Impact on Perceived Audio-Visual Quality in Different Interactive Contexts. In *Proc. IEEE Int. Workshop on Multimedia Signal Processing (MMSP'09)*.

Benoît, C., Martin, J.-C., Pelachaud, C., Schomaker, L., and Suhm, B. (2000). Audio-Visual and Multimodal Speech-Based Systems. In Gibbon, D., Mertins, I., and Moore, R. K., editors, *Handbook of Multimodal and Spoken Dialogue Systems*, pages 102–203. Kluwer Academic Publ., Boston MA.

Coutaz, J., Nigay, L., Salber, D., Blandford, A., May, J., and Young, R. (1995). Four Easy Pieces for Assessing the Usability of Multimodal Interaction: The CARE Properties. In Nordby, K., Helmersen, P., Gilmore, D., and Arnesen, S., editors, *Human-Computer Interaction, Interact '95*, pages 115–120. Chapman & Hall, London.

Dehn, D. M. and Van Mulken, S. (2000). The Impact of Animated Interface Agents: a Review of Empirical Research. *International Journal of Human-Computer Studies*, 52(1):1–22.

Garcia, M. and Raake, A. (2009). Impairment-Factor-based Audio-Visual Quality Model for IPTV. In *Proceedings of the 1st International Workshop on Quality of Multimedia Experience (QoMEX'09)*.

Hassenzahl, M., Burmester, M., and Koller, F. (2003). AttrakDiff: Ein Fragebogen zur Messung wahrge-nommener hedonischer und pragmatischer

Qualität. In Ziegler, J. and Szwillus, G., editors, *Interaktion in Bewegung. Proc. Mensch & Computer '03*, pages 187–196, Stuttgart. B.G. Teubner.

ITU-T Rec. E.800 (1994). Terms and Definitions Related to Quality of Service and Network Performance Including Dependability. International Telecommunication Union, Geneva.

ITU-T Rec. G.107 (2005). *The E-model, a Computational Model for Use in Transmission Planning*. International Telecommunication Union, Geneva.

ITU-T Rec. P.10 (2007). Vocabulary for Performance and Quality of Service. International Telecommunication Union, Geneva.

ITU-T Rec. P.851 (2003). Subjective Quality Evaluation of Telephone Services Based on Spoken Dialogue Systems. International Telecommunication Union, Geneva.

ITU-T Rec. P.920 (2000). *Interactive Test Methods for Audiovisual Communication*. International Telecommunication Union, Geneva.

Jekosch, U. (2004). Basic Concepts and Terms of "Quality", Reconsidered in the Context of Product-Sound Quality. *Acta Acustica united with Acustica*, 90(6):999–1006.

Jekosch, U. (2005). *Voice and Speech Quality Perception. Assessment and Evaluation*. Springer, Berlin.

Kamii, C., Lewis, B. A., and Kirkland, L. D. (2001). Subtraction Compared with Addition. *Mathematical Behavior*, 20:33–42.

Kühnel, C., Weiss, B., and Möller, S. (2009). Talking Heads for Interacting with Spoken Dialog Smart-Home Systems. In *Proceedings of the 10th Ann. Conference of the Int. Speech Communication Assoc. (Interspeech'09)*, pages 304–307.

Kühnel, C., Weiss, B., Wechsung, I., Fagel, S., and Möller, S. (2008). Evaluating Talking Heads for Smart Home Systems. In *Proceedings of International Conference on Multimodal Interfaces (ICMI'08)*.

López-Cózar Delgado, R. and Araki, M. (2005). *Spoken, Multilingual and Multimodal Dialogue Systems: Development and Assessment*. John Wiley & Sons, Chichester.

Möller, S. (2000). *Assessment and Prediction of Speech Quality in Telecommunications*. Kluwer Academic Publishers, Boston.

Möller, S. (2005). *Quality of Telephone-based Spoken Dialogue Systems*. Springer, New York.

Möller, S., Engelbrecht, K.-P., Kühnel, C., Wechsung, I., and Weiss, B. (2009). A Taxonomy of Quality of Service and Quality of Experience of Multimodal Human-Machine Interaction. In *Proceedings of the 1st Internationl Workshop on Quality of Multimedia Experience (QoMEX'09)*.

Möller, S., Engelbrecht, K.-P., Kühnel, C., Wechsung, I., and Weiss, B. (2010). Evaluation of Multimodal Interfaces for Ambient Intelligence. In Aghajan,

H., López-Cózar Delgado, R., and Augusto, J. C., editors, *Human-Centric Interfaces for Ambient Intelligence*, pages 347–370. Elsevier, Amsterdam.

Oviatt, S. (2004). Multimodal Interfaces. In Sears, A. and Jacko, J., editors, *The Human Computer Interaction Handbook*, pages 413–432. Lawrence Erlbaum, New York, 2 edition.

Oviatt, S., Coulston, R., and Lunsford, R. (2004). When do we Interact Multimodally? Cognitive Load and Multimodal Communication Patterns. In *Proceedings of the 6th International Conference on Multimodal Interfaces*, pages 129–136.

Sarter, N. (1995). Multiple-Resource Theory as a Basis for Multimodal Interface Design: Success Stories, Qualifications, and Research Needs. In Kramer, A., Wiegmann, D., and Kirlik, A., editors, *Attention: From Theory to Practice*, pages 187–195. Oxford University Press.

Schomaker, L., Nijtmans, J., Camurri, A., Lavagetto, F., Morasso, P., Benoît, C., Guiard-Marigny, T., Le Goff, B., Robert-Ribes, J., Adjoudani, A., Defée, I., Münch, S., Hartung, K., and Blauert, J. (1995). *A Taxonomy of Multimodal Interaction in the Human Information Processing System*. NICI, Nijmegen.

Wechsung, I., Engelbrecht, K.-P., Nauman, A., Schaffer, S., Seebode, J., Metze, F., and Möller, S. (2009a). Predicting the Quality of Multimodal Systems Based on Judgements of Single Modalities. In *Proceedings of the 10th Ann. Conf. of the Int. Speech Communication Assoc. (Interspeech'09)*, pages 1827–1830.

Wechsung, I., Engelbrecht, K.-P., Schaffer, S., Seebode, J., Metze, F., and Möller, S. (2009b). Usability-Evaluation multimodaler Schnittstellen: Ist das Ganze die Summe seiner Teile? In Kain, S. and Struve, D., editors, *Grenzenlos frei. Proc. Mensch & Computer '09*, pages 495–498. Oldenbourg Wissenschaftsverlag.

Weiss, B., Kühnel, C., Wechsung, I., Möller, S., and Fagel, S. (2009). Comparison of Different Talking Heads in Non-Interactive Settings. In *Proceedings of Human Computer Interaction International (HCII), San Diego*, pages 349–357.

Weiss, B.and Kühnel, C., Wechsung, I., Fagel, S., and Möller, S. (2010). Quality of Talking Heads in Different Interaction and Media Contexts. *Speech Communication*, 52(6):481–492.

Wickens, C. (1999). Multiple Resources and Performance Prediction. *Theoretical Issues in Ergonomics Science*, 3:159–177.