EURASIP Journal on
Audio, Speech, and Music Processing
a SpringerOpen Journal

RESEARCH Open Access

Audiovisual quality integration for interactive communications

Benjamin Belmudez[*] and Sebastian Möller

Abstract

This paper investigates multi-modal aspects of audiovisual quality assessment for interactive communication services. It shows how perceived auditory and visual qualities integrate to an overall audiovisual quality perception in different experimental contexts. Two audiovisual experiments are presented and provide experimental data for the present analysis. First, two experimental contexts are compared, i.e., passive 'viewing and listening' and interactive, with regard to their impact on the audiovisual qualities as subjectively perceived by the user. Second, the effects of cross-modal interactions on the assessment of the audio and video qualities are measured for those experimental contexts. The results are compared to the ones found in the literature revealing both similarities and differences in terms of magnitude and also in which cases they occur. Third, the impact of the conversational scenario on the assessment of the auditory and visual qualities is investigated. Finally, results from the literature related to audiovisual integration are gathered by the type of application. A general integration function is proposed for each category, and the performances of these 'application-oriented' models demonstrate a direct gain in prediction.

Keywords: Interpersonal communication; Audiovisual quality integration; Cross-modal interaction; Subjective quality assessment; Quality modeling

1 Introduction

Audiovisual telecommunication services like video on demand (VoD), Internet protocol television (IPTV), mobile television (MoTV), and videotelephony became broadly used multimedia services in the past two decades. In order to ensure a high level of user satisfaction, it is required to efficiently assess the quality of experience (QoE) with regard to the quality of service (QoS) through the variety of audiovisual services. The present study focuses on interactive video services like videotelephony and proposes to evaluate the audiovisual quality as subjectively perceived by the user. Perceived audiovisual quality stems from the cognitive processing of multi-sensory stimuli like the auditory and visual ones. In order to predict the quality of the perceived auditory and visual signals, instrumental models were developed. They are categorized depending on their inputs: parameter-based models use parameters from the application and the network layers; media-based models perform a quality analysis on the physical characteristics of the transmitted signal itself. Finally, hybrid models constitute a mixture of the media-based and bitstream-based models. For further information on the current state of the art on audio and video quality metrics, one can refer to exhaustive surveys on that particular topic [1-3]. The aforementioned models were developed in a passive paradigm and assume a single modality (audio or video).

Several aspects pertaining to human perception should be considered when building a multi-sensory perceptual model. Hollier et al. proposed such a theoretical model that includes the degradations on the sensory streams, cross-modal interactions (inter-stream synchronization and cross-modal influences on quality), as well as a task-related perceptual layer [4]. This layer can refer to the mode of communication (passive watching, interactive videotelephony, remote teaching, etc.) or the task itself (listening to a speech, looking at documents, etc.). Its architecture as well as its level of granularity are elements which were not explicitly defined as they depend on the type of application. The impact of the communication mode on multi-modal quality can be evaluated by adapting the experimental context. This latter will here refer

[*]Correspondence: ben.belmudez@gmail.com
Quality and Usability Lab, Telekom Innovation Laboratories, Technische Universität Berlin, Ernst-Reuter-Platz 7, Berlin 10587, Germany

to the degree of interactivity, if any, experienced by the subjects during the quality evaluation phase.

Concerning cross-modal interactions, experimental evidence support the hypothesis that sensory streams actually interact: signals proceeding from different modalities are not processed separately [5]. For instance, the well-known 'McGurk' effect illustrates this distinctive characteristic of the human information processing system [6]. It shows that the human perception of speech is bimodal as the acoustic speech can be affected by visual cues from lip movements. The characteristics of these cross-modal interactions were studied in a passive experimental context, and literature reports heterogenous results: in particular, the type of content was shown to be an influencing factor. Strong differences arose when comparing cross-modal effects between virtual reality content (fly-through with narration) and 'talking head' content [4]. In the first case, no inter-modal dependency could be observed as in the second case, both auditory and visual modes interacted in a significant way. For television contents (e.g., reports, news, sport), adding a high-quality soundtrack could either lessen the perception of the visual impairments or combine to provide an impression of better visual quality [7]. Due to the bimodality in speech perception, audiovisual interaction is an important design factor for multi-modal communication systems such as videotelephony. It is however unclear whether cross-modal effects do have an impact on the audiovisual integration and if they should be accounted for in the construction of a perceptual audiovisual model.

Transmission delay can have a significant influence on the interactivity of a conversation and thus on the perceived quality. Early studies on the impact of conversational delay on speech quality, like the one from Kitawaki and Itoh [8], predicted a severe impact of transmission delay on the perceived speech quality especially when considering highly interactive tasks like the 'Random Number Verification' task [9]. However, recent studies [10-12] suggested that the impact of delay in close-to-natural conversation scenarios is far less important as suggested in [8]. In the case of everyday life conversations, it was observed that people tend to adapt their communication strategy affecting both the structure and the semantic flow of the conversation (e.g., misunderstandings and interruptions). Still opened is the question about the relative contribution of pure delay impairment to other types of degradations such as coding or network packet loss. In Yamagishi and Hayashi [13], the authors addressed the audiovisual quality prediction of interactive multimedia services such as videotelephony through the use of perceptual dimensions. They identified two dimensions, namely 'aesthetic feeling' and 'feeling of activity.' The first dimension referred to factors like audio and video coding and packet loss, namely factors distorting the shape of the audiovisual signal. The second dimension was affected by factors related to the temporal continuity like one-way transmission delay and video frame rate. In their experiment, they used the 'Name Guessing Task' [9] conversational scenario which has a relatively low interactivity level. In that case, they found the aesthetic feeling dimension to carry more weight in the overall quality perception. However, other findings indicate that for highly interactive scenarios, the impact due to delay predominates quality perception. In contrast, Gros et al. studied the impact of three factors on conversational speech quality: temporal quality degradation profile, network packet loss (between 0% and 30%), and jitter (between 500 and 800 ms) [14]. They used short conversation test (SCT) scenarios defined in [15] which are close-to-natural conversational scenarios (e.g., ordering a pizza) and exhibit a middle to low level of interactivity. They pointed out that subjects were sensitive to variations in the transmission characteristics (temporal quality profile) and to packet loss. In turn, adding jitter to packet loss would not lead to a significant decrease in quality ratings. Therefore, delay as an impairment factor may belong to a different perceptual quality dimension, and its weight on the overall quality evaluation compared to other types of impairments like coding and network packet loss impairments depends on the interactivity induced by the conversational scenario. In this study, the focus is brought into the influence of the combined effect of coding and transmission factors and of the type conversational scenario on the perceived audiovisual conversational quality.

Finally, audiovisual quality integration in the case of multi-modal services defines how users process multiple sources of information and aggregate these information (or derived attributes) into an overall perceived quality impression. Descriptive studies, encompassing different kinds of applications, converge to agree that audiovisual quality can be predicted on the basis of the individual auditory and visual qualities. Nevertheless, significant gaps in knowledge remain concerning the human perceptual processing of sensory information, as no neurophysiological model came to describe the binding between different functional areas of the brain and thus multimodal processing [16]. Current models for audiovisual quality prediction were developed and trained for specific applications (i.e., IPTV, mobile streaming, etc.) and communication modes. Hence, there is a need for understanding which factors actually do impact the audiovisual quality integration, positing that it can be accurately predicted from single modalities.

In this paper, we investigate several aspects of audiovisual perception within the framework of interactive video communication services. First, we will compare two different experimental contexts: passive viewing and listening versus interactive. Second, we take a closer look at the cross-modal interactions occurring for both experimental contexts with regard to the impact of cross-modal effects on the audiovisual integration. Third, focusing on the interactive situation, we will quantify the influence of the conversational scenarios on the audio, video, and audiovisual qualities. Finally, we propose four general audiovisual integration functions for specific applications jointly based on the results presented in this analysis and on the results from the literature.

2 Related work

2.1 Audiovisual quality perception

The psychophysical processes involved in the perception of uni-modal stimuli (e.g., visual or auditory) have been well established. Audiovisual perception is a multi-modal process that consists of the integration of both visual and auditory sensory channels. This multi-modal processing of information suffers from a lack of understanding from a neurophysiological point of view: how do the neurons of the specific cortical areas for single sensory perception communicate? In summary, how is the information coming from different functional areas shared in order to achieve multi-modal processing, and at what stage of the processing do cross-modal interactions occur? [16]

Even though the low-level processing details remain unknown, there is empirical evidence demonstrating certain key characteristics of the multi-modal perception, for example, information coming from one sensory modality can be influenced by information coming from another sensory modality (inter-sensory biases). The different sources of information are not processed independently, but they are integrated: new information is produced that could not have been obtained with a single modality. Cross-modal studies showed several implications of this characteristic: modalities can influence each other not only on thresholds (e.g., ability to detect visual motion influenced by sound) but also on the intensity of the perception itself, when one modality improves the experience of another modality [17]. For instance, under impaired hearing conditions (e.g., background noise), speech intelligibility can be greatly enhanced by adding a visual channel showing the lip movements of the speaker's mouth [18]. A spatial and temporal lack of proximity between modalities can impair the bimodal integration [19], e.g., the ventriloquist effect describes the effect of sound source perception modified by a visual stimulation. Inter-stream asynchrony can also hinder the perception, particularly in the case of videotelephony where the lips of the speaker are clearly visible and with audio led asynchrony.

According to the modality appropriateness hypothesis [20], the more suitable sensory modality (e.g., in terms of accuracy like spatial or temporal resolution for a given task) will tend to have a stronger influence on the multimodal perception. It would stem from the differences in the suitability of the modalities for the perceptual coding of a certain stimulus feature [21]. It has been hypothesized that the visual modality can be dominant for spatial task and the audio modality for temporal ones due to their respective resolution accuracy [22]. A task dependency was found when investigating the relative importance between audio and visual information with respect to the interaction scenario (human-human or human-machine interaction) and to the degree of interactivity [23,24].

2.2 Impact of the experimental context on perceived audiovisual quality

The evaluation of human perception of audiovisual quality also depends on the employed experimental methodology. In particular, the situation of assessment in which the judging subject is placed (listening/viewing or conversational context referred as experimental context) can impact the judgement process [11]. Subjective tests should reflect the ecological environment of the service or application under assessment. For evaluating conversational speech quality, conversation tests constitute a realistic situation where a natural behavior can be expected from test participants. This interactive situation differs from a passive one as the assessment task, in a passive paradigm, is conducted without any other cognitive load than the one caused by watching or listening to the stimuli. The interactive situation impacts the perception mainly because the attentional resources are split between the task of assessing the quality and the activity of communication [14]. It was hypothesized by Kahneman that the attentional resources are limited in quantity [25]. Therefore, the sharing of attention between two tasks can potentially hinder the cognitive processes of either integrating the quality or evaluating it. Indeed, in the case where interactants firstly focus on the content of what is said or viewed, less attentional resources will be dedicated to analyzing the form of the auditory and visual signals thus leading to fewer diagnostic information describing these signals [15]. As a result, quality judgements in an interactive situation could diverge from those obtained in a passive situation of assessment. For that reason, Hollier et al. [4] mentioned the need to take the granularity of the task (passive watching, one-to-one conversation, etc.) into account within the process of building a multi-modal model for subjective quality prediction.

A study from Gros et al. [14] on the impact of the experimental context on speech quality reported that subjective judgments were similar between the listening and the conversational contexts. It was stated that the 'conversation doesn't seem to disturb the perception, integration and the memorization (*cf.* recency effect) of the degradations and their variations, nor the elaboration of a quality judgment. However the range of judgments in the conversational context appeared to be more limited than in a listening situation.' For the assessment of audiovisual quality, Chateau [26] compared passive and interactive contexts using 10-s videoconferencing clips for the passive context and one interactive scenario (similar to the Name Guessing Task described in the ITU-T Rec. P.920 [9]) for the interactive context. They reported similar video quality ratings for both contexts, but the MOS range of the audio scores was significantly reduced for the interactive situation. A possible explanation was that audio was rather judged in terms of acceptability in the interactive situation. The difference when comparing an interactive to a passive context of assessment could be the loss of discrimination (reduced MOS range) and potentially an asymmetrical assessment depending on the modality that interactants dedicate most of their shared attentional resources to.

2.3 Cross-modal interactions

Cross-modal perception involves interactions between two or more different sensory modalities. Empirical observations showed that one modality can modify the perceptual experience formed by another modality. Quality experiments involving different audiovisual contents and communication modes reported heterogenous results. As stated earlier, when video accompanies the acoustic utterance, it increases the speech intelligibility, thanks to the visual information brought by the lips' movements [5]. For the passive evaluation of videotelephony content (head-and-shoulders with a fixed background), a study from Rimmel et al. revealed strong mutual compensation between modalities [27]. Increasing the quality of one modality significantly improved the perceived quality of the other modality. This experiment was based on the evaluation of 6-s video clips consisting of a talker's upper body (two males and two females). A similar study from Chateau, using 10-s video clips of videotelephony material (one male and one female), did not demonstrate any influence of the audio channel on the perceived video quality and only a weak influence of the video channel on the perceived audio quality [26]. The fact that the audio quality levels used in that experiment were above the intelligibility level could explain that the perception of video quality was independent from the audio quality level. Within the same study, cross-modal interactions were investigated for an interactive context where a pair of interactants had to carry out a conversational task (Name Guessing Task) through an audiovisual link established between two separate rooms. Results between both passive and interactive contexts were similar except for the weak effect of the video channel on the perceived audio quality found in the passive context that became more conspicuous in the interactive context.

Such contradicting results were also found for experiments using television material. Two studies support the hypothesis that television images presented with a high-quality soundtrack are more 'involving' and of better quality [7,17]. However, another study from Beerends [28] based on 25-s commercials reported asymmetric interaction effects with a noticeable influence of the video quality level on the perceived audio quality (0.5 on a 5-point MOS scale) and a weaker influence of the audio quality level on the perceived video quality (only 0.15 on a MOS). Comparing these results to the ones obtained with head-and-shoulders material, Hands pointed out that the nature of the audiovisual content may have influenced the results as commercials are visually more captivating, thus leading to a more video dominant situation [29].

In a recent survey on audiovisual quality assessment [3], the authors concluded that 'when measuring individual audio or video quality in audio-visual stimuli, the influence of the other modality might be small, but cannot be neglected totally.' It is yet unclear if this mutual influence also has an impact on the audiovisual integration. Even though cross-modal interactions are reported in the aforementioned studies, the presence and the magnitude of these effects strongly depend on the audiovisual content and on the experimental context.

2.4 Audiovisual quality integration

As stated in Section 1, fairly accurate quality metrics were developed for the audio and video modalities. They are based on the comprehension of the psychophysical processes involved in the auditory and visual perception. In audiovisual perception, it remains undetermined at which stage of the perceptual processing chain the modalities do actually combine. Therefore, there is yet no clear cognitive basis that explicitly describes how users of multimedia services integrate information from different sources (audio, video, haptic, etc.) to form an overall quality judgement [16]. Researchers have turned toward theories of attention as an attempt to bring some insight into the audiovisual perception process. The preferred theory, called late fusion, states that the auditory and visual signals are internally processed to produce separate auditory and visual qualities that are fused at a late stage to give a judgment of the overall perceived quality [3]. Audiovisual quality is thus generally described as a combination of two dimensions (audio and video qualities) leading to the following integration model:

$$\mathrm{MOS_{AV}} = \alpha \mathrm{MOS_A} + \beta \mathrm{MOS_V} + \gamma \mathrm{MOS_A} \cdot \mathrm{MOS_V} + \zeta, \tag{1}$$

with $\mathrm{MOS_{AV}}$ being the audiovisual quality; $\mathrm{MOS_A}$, the audio quality; $\mathrm{MOS_V}$, the video quality; and α, β, γ, and ζ, the scalar coefficients. Several early experiments were conducted in the 90s to derive this mathematical formula that performs the audiovisual integration from the quality metrics of the single modalities

[26,30-33]. Based on the results of those experiments, the International Telecommunication Union (ITU) proposed to only use the multiplicative term between the audio and video qualities ($\mathrm{MOS_A} \cdot \mathrm{MOS_V}$) with an additive shift as an estimator of the audiovisual quality [34]. The recommended values of the equation coefficients are an average of the values derived in the studies mentioned above. Further experiments were conducted in order to derive audiovisual integration functions fitted for specific applications like mobile television [35,36] or high-definition television [37]. Still, as stated by You [3], 'there is no reliable metric available for measuring the audio-visual quality automatically.'

Indeed, the results reported in the literature were derived from experiments conducted in various setups depending on the targeted applications and following different testing methodologies, test conditions (range and type of audiovisual impairments), audiovisual stimuli and presentation devices. The resulting integration coefficients were different between these experiments and were usually optimized over one dataset, therefore not directly applicable to other cases. A meta-analysis performed by Pinson et al. [38] compared on a high level these integration models and concluded that the MOS ranges of video and audio qualities are of primary importance: it was hypothesized that if the variation in MOS range for one modality is significantly greater than the other, it may introduce a bias resulting in having one modality appearing to be more correlated with the audiovisual quality. This analytical consideration refers to the test design and how an unbalanced experiment can lead to biased conclusions concerning the relative importance of the audio and video modalities in the human information process.

3 Research method

3.1 Test procedure

We carried out two subjective audiovisual experiments. In the first experiment, the quality of short samples was evaluated in a passive context (listening and viewing only). The content of these samples was a 'head-and-shoulders' videotelephony scene with a fixed background. In each sample, a speaker uttered one or two sentences. Each sample was about 9-s long in order to be in line with ITU-T Rec. P.911 for the subjective evaluation of audiovisual stimuli. After each video clip, test participants were asked to rate the audiovisual, video, and audio qualities in that specific order. To that end, they used the 11-point continuous rating scale defined in ITU-T Rec. P.910 [39]. The stimuli were presented in a randomized order so that two consecutive sequences could not be of the same speaker or have the same test condition.

In the second experiment, pairs of participants were invited to carry out a set of interactive videotelephony conversations using different conversational scenarios. For each conversation, participants were provided with a description of the scenario and a set of instructions as guidance for the realization of the conversation. The conversing partners were informed beforehand about the content of the scenario in order for them to run the conversation smoothly and in a natural way. At the end of each conversation, test participants had to rate the audiovisual, video, and audio qualities of the entire conversation using the same 11-point scale as in the first experiment. We split the test participants into three groups of eight pairs, leading to 24 ratings per test condition. Each pair of participants experienced all conditions but only associated with a subset of all possible scenarios. That distribution was designed to keep the total duration of the experiment under 1 h. The quality assessment methodology followed the procedure described in the ITU-T Rec. P.920 for the assessment of audiovisual interactive communications.

The experimental procedure involved in this study does comply with the Helsinki declaration and has been discussed with and approved by the ethics committee of the Technische UniversitŁt Berlin.

3.2 Test participants

Twenty-one naïve participants carried out the first experiment, and 24 pairs of participants carried out the second experiment. Prior to the experiments, participants were screened for normal visual acuity and color vision according to ITU-T Rec. P.911. They were not experienced assessors and were not involved in image or audio quality evaluation as part of their work. All of them received a monetary compensation.

3.3 Test bed

For the passive experiment, the test sequences were displayed on a 10.1-in. ($1,024 \times 600$ pixels) laptop screen and on a 19-in. LCD monitor with a resolution of $1,080 \times 1,024$ pixels for the interactive experiment. The audio playback was realized using a high-quality sound card (Edirol UA-25, Roland Corp., Los Angeles, CA, USA) and headphones (Sennheiser HMD 410, Hanover, Germany). The test participants were sitting in a room designed for video testing and compliant with the experimental listening and viewing conditions defined in ITU-T Rec. P.911. The viewing distance was approximately 30 cm for the passive experiment and 50 cm for the interactive one, both corresponding to three times the height of the picture. Both distances were within the range recommended in ITU-T Rec. P.911 and ensured a viewing angle (apparent size of the image) identical between both experiments.

A videotelephony client [40] was developed based on the Voice over IP framework PJPROJECT 0.8.3 [41]. This client could be used in two modes: (1) processing mode ('off-line'), where the original (unimpaired) audiovisual sequences were processed prior to the experiment and according to the test conditions. This mode was used to produce the stimuli for the passive experiment. (2) Videotelephony mode ('on-line'), where an audiovisual link between two experiment rooms was established. The audio and video quality parameters of the multimedia link were controlled separately. This mode was used for the interactive experiment.

For the passive experiment, stimuli were displayed using an interface with a uniform gray background. After each stimulus, rating scales were displayed on the screen for the quality evaluation phase. For the needs of the interactive experiment, a second user interface was designed to guide the participants through the test, allowing them to be autonomous during the test session, i.e., controlling when the conversation should start and end. Rating scales for the evaluation phase were automatically displayed after each conversation.

3.4 Audiovisual stimuli

3.4.1 *Passive experiment*

Audiovisual recordings of two German speakers (one male and one female) were realized using different scene backgrounds and conversation topics. The stimuli were recorded in raw format (uncompressed planar YUV 4:2:0) with a VGA (640 × 480) resolution and a frame rate of 25 frames per second. The audio recordings were made using a sampling frequency of 16 kHz and 8-bit quantization. Once recorded, the stimuli were degraded to achieve seven equidistant targeted levels of quality for both the audio and video channels leading to 49 possible predefined combinations of audio and video quality levels for each sample. From these combinations, 33 were selected for testing. Each quality level depends on a set of selected parameters related to the application and the network. For video, these parameters were the type of video codec (H.264 codec taken from [42]), the video encoding bit rate (Br), and information loss introduced by dropping data packets of streamed video with a fixed packet loss (Pl) rate. Packet loss was achieved through a network emulator (*netem* module [43]) with a random loss distribution. The parameters' values have been determined using a pretest containing seven different test conditions. The average MOS values were found to span the entire quality range. For audio, we selected two speech codecs: a narrowband codec (GSM-EFR) and a wideband codec (AMR-WB). The WB-E-model [44], which is a planning model for the quality of transmitted speech, was used to help determine the values of network Pl that would lead to the targeted levels of QoE. The same packet loss emulation tool was used for video and audio. The relevant QoE parameters of the audio and video channels which were varied in the experiments are presented in Table 1. The chosen values of the parameters for each quality level are listed in Table 2. Lv1 and La1 denote levels of poor quality, whereas Lv7 and La7 denote levels of high quality for video and audio, respectively.

Table 1 Details of the experimental conditions for the audiovisual experiments

Context name	Passive Exp. 1	Interactive Exp. 2
Video codecs and operating bit rates (Mbps)	H.264 at {0.4, 0.77, 2}	H.264 at {0.4, 0.77, 2}
Audio codecs and operating bit rates (kbps)	WB: G.722.2 at 23.05, NB: GSM-EFR at 12.2	WB: G.722 at 64, NB: GSM-EFR at 12.2
Video packet loss rate (%)	0, 0.5, 3, 5	0, 3, 5
Audio packet loss rate (%)	0, 3, 7, 20	0, 3, 20
Video contents	2 clips H&S	-
Interactive scenario	-	SCT, AVSCT, BB
Number of test conditions	33	16

Subjects	$N = 21$ (9 m, 12 w),
	mean = 25.9 years,
	std = 4.1 years
		$N = 48$ (22 m, 26 w),
		mean = 27.3 years,
		std = 5.6 years

Exp. 1, experiment 1; Exp. 2, experiment 2; WB, wideband; NB, narrowband; std, standard deviation; m, men; w, women.

3.4.2 *Interactive experiment*

Three different conversational scenarios were used in the interactive experiment. The first one is referred as 'SCT' which stands for short conversation test. It has been developed for use in audio-only conversations [15] and involves mainly the audio channel. The second type of scenario called audiovisual short conversation test ('AVSCT') can be considered as an audiovisual version of the SCT scenario in the sense that visual cues were added to the dialogs. This scenario is intended to simulate an 'average' videotelephony conversation with a balanced use of the audio and video channels. It consists of a semi-structured dialog where interactants alternately answer each other's questions. These dialogues have been developed for the German language. An extract of the 'car renting' scenario is provided below:

- Interlocutor 1: *Was für ein Auto möchten sie mieten?* (What kind of vehicle would you like to rent?)
- Interlocutor 2: *Einen Kombi, zeige Bild eines Kombis* (A break, show a picture of a break). *Wie sieht der Kleintransporter im Angebot aus?* (How does the small pick-up look like in the offer?)
- Interlocutor 1: *Beschreibe den Kleintransporter (Farbe...) und zeige ein Foto* (Describe the small pick-up (color etc.) and show a picture). *Wann wollen sie das Fahrzeug mieten?* (When do you want to rent the vehicle?)

Table 2 Description of the audio and quality levels used for the audiovisual experiments

Video quality levels		Exp. 1	Exp. 2	Audio quality levels		Exp. 1	Exp. 2
Lv1	2 at 5%	x	x	La1	NB at 20%	x	x
Lv2	0.4 at 3%	x	x	La2	WB at 20%	x	x
Lv3	2 at 0.5%	x	-	La3	NB at 4%	x	-
Lv4	0.4 at 0%	x	-	La4	NB at 0%	x	x
Lv5	0.77 at 0.5%	x	-	La5	WB at 7%	x	-
Lv6	0.77 at 0%	x	x	La6	WB at 3%	x	x
Lv7	2 at 0%	x	x	La7	WB at 0%	x	x

A full description of this scenario and its implementations can be found in [45]. Finally, the third scenario is the building block scenario ('BB') described in ITU-T Rec. P.920, where the use of the video channel was found to be predominant [45]. The participants were asked to perform the tasks associated with each scenario as efficiently as possible in order to limit the duration of each conversation. These were intended to last around 3 min. We selected four levels of video quality combined with three levels of wideband audio quality codec leading to 12 audiovisual conditions. Additionally, 4 conditions with the narrowband codec were selected that resulted in a total of 16 test conditions.

4 Influence of the experimental context on the audiovisual modalities

In this section, we will look at the influence of the experimental context (passive vs. interactive) on audiovisual quality. We will perform a cross-experimental comparison of the quality ratings (audio, video, or audiovisual) obtained for similar test conditions between both experiments.

4.1 Influence of the experimental context on subjective video quality

The subjective video quality scores obtained for different video quality levels and for two different experimental contexts are represented in Figure 1. A univariate analysis of variance computed on the video MOS revealed a main effect of the video quality level ($F(6, 2073) = 328.5, p < 0.01$) and of the audio quality level ($F(6, 2073) = 4.4, p < 0.01$) and a significant influence of the experimental context ($F(1, 2073) = 13.1, p < 0.01$). No interaction effect was detected between the factors 'audio quality level' and 'experimental context.' The impact of the experimental context on the video MOS did not depend on the level of audio quality. The video scores displayed in Figure 1 were thus computed over all audio quality conditions.

For the passive experiment, video levels were different from each other according to a *post hoc* test (Scheffé, $p < 0.05$), except for levels Lv3 and Lv4, and levels Lv4 and Lv5. For the interactive experiment, only Lv6 did not reach the level of statistical difference when compared to Lv2 and Lv7.

In a passive context, test participants can dedicate their attentional resources to assessing multimedia quality. They are therefore expected to be more discriminant in detecting visual impairments and therefore conscious distinctions between different levels of quality can be established. In an interactive context, they are asked to fulfil a task which requires attentional resources. In the latter case, quality ratings were clustered on a smaller portion of the scale. Our interactive tasks (referred as conversational scenarios) drove the test participants to avoid using the low end of the scale. Hence, when presented with identical levels of poor video quality, test participants judged the quality significantly worse in a passive context than in an interactive context. However, ratings collected for levels of high video quality did not differ between both experimental contexts.

4.2 Influence of the experimental context on subjective audio quality

A univariate analysis of variance performed on the audio MOS revealed a main effect of the audio quality level ($F(6, 2073) = 209.5$, $p < 0.01$) and the video quality level ($F(6, 2073) = 5.7$, $p < 0.01$) and a significant impact of the experimental context ($F(1, 512) = 54$, $p < 0.01$). An interaction effect was detected between the factors 'video quality level' and 'experimental context.' The impact of the experimental context on the video MOS did depend on the level of video quality ($F(3, 2073) = 3.4$, $p < 0.05$). However, this effect was weak in comparison to the impact of the experiment and only affected the differences in audio scores between both experimental contexts toward low audio quality levels (La1 and La2). It however did not change the general effect of the experiment. The audio scores presented in Figure 2 were thus computed over all video conditions.

For the passive context, several audio quality levels did not significantly differ according to a *post hoc* test: levels La2, La3, and La4 when compared to each other, and La4 compared to La5 (Scheffé, $p > 0.05$). For the interactive context, all audio levels were different

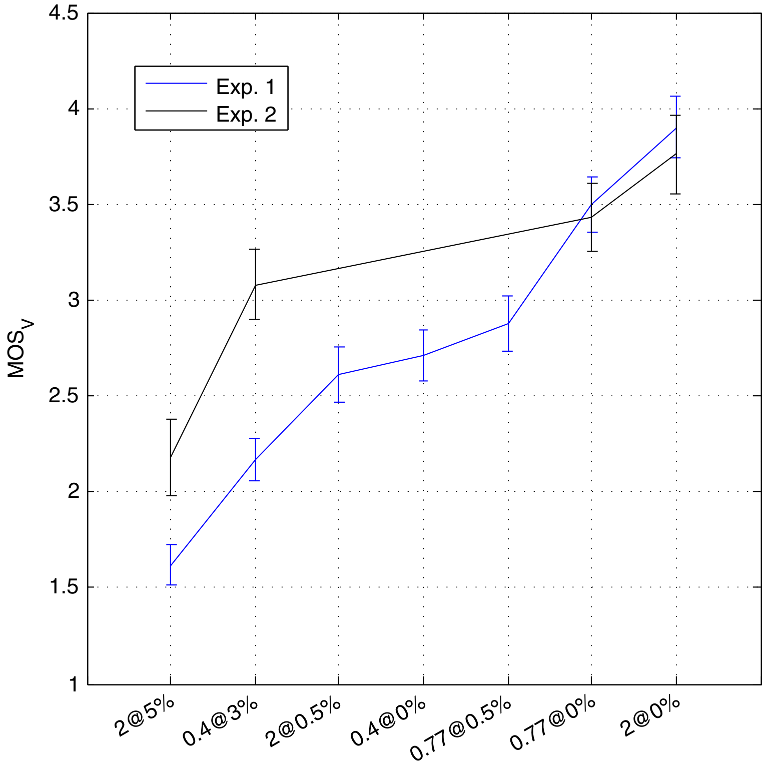

Figure 1 Effect of the experimental context on subjective video quality. Scatter plot of the subjective video scores with their 95% confidence intervals obtained for different levels of video quality in Exp. 1 and Exp. 2.

from each other except for levels Lv4 and Lv6 (Scheffé, $p > 0.05$).

As can be seen in Figure 2, the impact of the experiment was significant for conditions of poor audio quality (levels La1 to La4) for which more optimistic ratings were given for the interactive context. The conditions of high audio quality were rated similarly for both experiments. The impact of the experimental context followed the same pattern for both visual and auditory modalities. It can be argued that the sole collection of quality ratings may not be sufficient for explaining the differences in ratings for our experimental contexts. For instance, a strong difference in quality ratings was observed for La4, which is a test condition including narrowband encoding with no packet loss, thus corresponding to an undisturbed telephone connection (no packet loss, echo, or delay). Why was this condition rated above the level of fairness (MOS = 3) in an interactive context and below in a passive context? Bearing in mind the goal of fulfilling the task, it is conceivable that test participants granted more importance to the criteria of intelligibility ('Is the speech quality good enough so that I can understand my interlocutor without any difficulty?') than to the criteria of quality (for example, dissociating narrowband from wideband encoding) when emitting the quality judgement. Indeed, the difference in subjective scores between conditions La4 and La7 (narrowband and wideband encoding with no packet loss) was much smaller in an interactive context than in a passive one.

4.3 Influence of the experimental context on subjective audiovisual quality

A univariate analysis of variance performed on the audiovisual MOS revealed a main effect of the interaction of the audio and video quality levels ($F(14, 1290) = 82.6$, $p < 0.01$) and a significant impact of the experiment ($F(1, 1290) = 75.2$, $p < 0.01$). Differences in audiovisual scores were spotted for low levels of video quality (Lv1 and Lv2) (see Figure 3). This observation is justified by the fact that both modalities present discrepancies between experimental contexts toward low quality levels. As audiovisual quality is the integration of both modalities, similar perceptual effects were observed. This experimental result also suggests that strong visual impairments impact less severely the audiovisual perception in an interactive context, implying a stronger influence of the video channel on the audiovisual quality in a passive context. The comparison of two different experimental contexts pointed out that quality judgements collected in an interactive context tend to be more optimistic toward low quality levels for

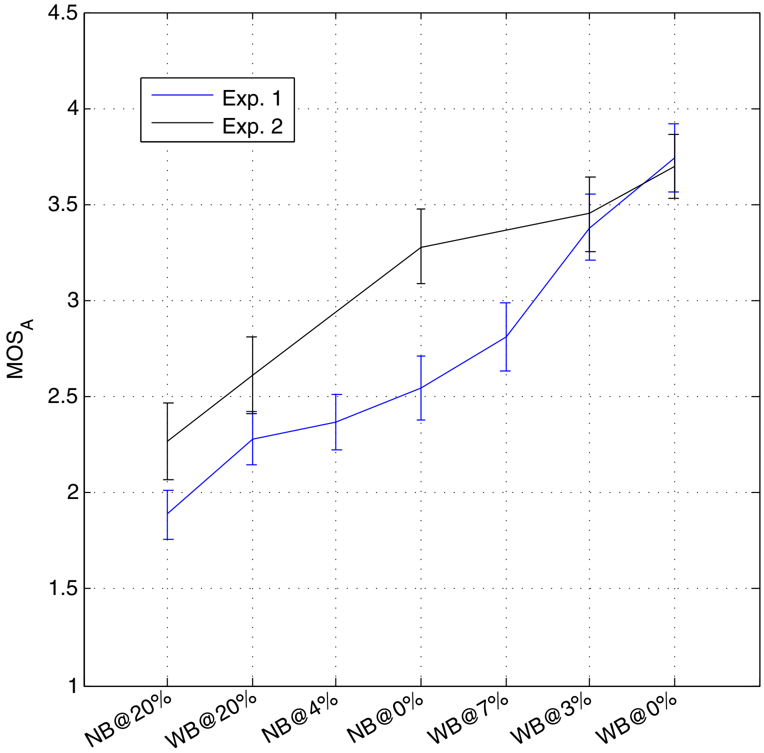

Figure 2 Effect of the experimental context on subjective audio quality. Scatter plot of the subjective audio scores with their 95% confidence intervals obtained for different levels of audio quality in Exp. 1 and Exp. 2.

all three modalities. The corresponding MOS ranges were therefore reduced due to poorer diagnostic information on the audio and video signals. The loss of discrimination occurring in an interactive context can be ascribed to the fact that attentional resources of the test participants were primarily dedicated to the realization of the task rather than to the analysis of the audiovisual content.

5 Influence of cross-modal interactions

In the last section, a statistical analysis of the subjective scores for both experimental contexts confirmed that the audio quality level has an influence on the video MOS and vice versa. In this section, cross-modal interactions will be investigated for each experimental context considered separately in order to determine for which particular test conditions these interactions actually occur. To complement this analysis, the type of conversational scenario will also be accounted for. Finally, the impact of cross-modal interactions on the audiovisual integration will be assessed for each experimental context.

5.1 Cross-modal interactions for passive testing

A straightforward method for representing cross-modal interactions and their impact on quality consists of comparing subjective scores obtained for one modality with a fixed quality level and vary, with a sufficient magnitude, the quality level of the other modality. In Figure 4, subjective scores for audio and video quality are represented for three distinct levels of quality for each modality (La1, La3, and La7 for audio and Lv1, Lv3, and Lv7 for video). The black vertical lines connect data points belonging to the same audio quality level and the horizontal lines connect data points belonging to the same video quality level. Lines of isoquality (dashed blue lines) were added in order to show where the data points should have theoretically been located in the absence of cross-modal interactions. The observed data clearly deviate from the isoquality lines demonstrating the quality impact of cross-modal interactions. An ANOVA was performed on the quality scores to detect the significant variations in ratings: a difference between the extreme levels of audio quality (La1 and La7) was found for the video level Lv3 and an almost significant difference for Lv7, with a maximal difference of 0.35 MOS. Similarly, the video quality level impacted the audio MOS: differences were detected for two levels of audio quality (La1 and La7) when comparing extreme levels of video quality (Lv1 and Lv7). The difference of video MOS for La7 was 0.54. A mutual influence of one modality onto the other was thus detected in both directions (of audio on video

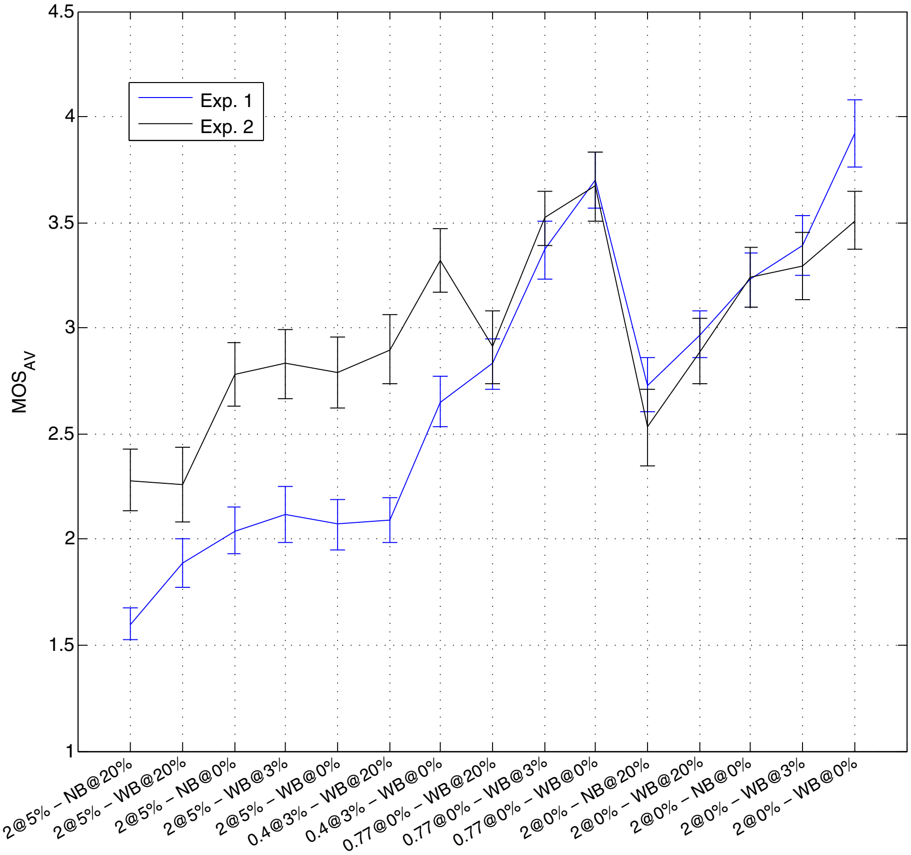

Figure 3 Effect of the experimental context on subjective audiovisual quality. Scatter plot of the subjective audiovisual scores with their 95% confidence intervals obtained for different combinations of video and audio levels and for Exp. 1 and Exp. 2.

and vice versa), but the differences in ratings were only statistically significant when comparing extreme levels of quality.

The second question under study consisted of determining whether the measured cross-modal interactions actually impacted the audiovisual quality integration and in such a case to which extent would it differ from an integration performed in the absence of cross-modal effects. To this end, we could compare the observed data to theoretical values taken as upper and lower bounds. The intersections of the dashed blue lines in Figure 4 represent the corrected theoretical locations of the data points in the space $\{MOS_A; MOS_V\}$ assuming the absence of cross-modal effects. There were two possibilities concerning the audiovisual scores that could be attributed to these points. The first hypothesis was that cross-modal effects modified the location of the points in the space $\{MOS_A; MOS_V\}$ but did not alter the audiovisual perception; thus, the same audiovisual score as their corresponding data points could be assigned. The second hypothesis was that if the location was modified, the audiovisual score should be computed accordingly. This score could be predicted on the basis of a non-linear multiple regression performed on the experimental dataset taking as a mathematical model of audiovisual integration the one described by Equation 1. The upper bound thus coincides with the predictive model trained on the experimental dataset.

Both hypotheses are illustrated in Figure 5, where the observed quality scores are assorted of their upper and lower bound values in the audiovisual space $\{MOS_A; MOS_V; MOS_{AV}\}$. A visual inspection of the data distribution suggests that there is not a great deal of variation between the lower and the upper bounds. The Pearson correlation computed between both bounds (0.98) and the low RMSE value (0.16) come to support this observation and confirm the tight proximity between both alternatives. Locations of noticeable disparity between both bounds are found in the corners of either high video quality combined with poor audio quality and the other way around, i.e., where cross-modal effects are statistically significant. However, the amplitude of variation at these locations is relatively modest and does not seem to impact the overall audiovisual integration.

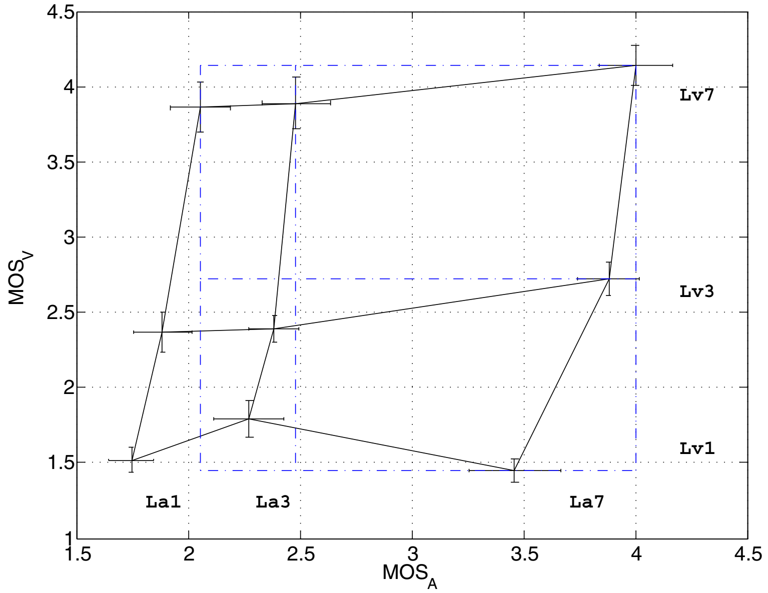

Figure 4 Cross-modal interactions for the passive experimental context (Exp. 1). Scatter plot of the subjective audio and video scores with their 95% confidence intervals obtained for different combinations of audio and video levels. Vertical and horizontal lines connect data points obtained for similar audio and video quality levels, respectively (La1, La3, and La7 for audio and Lv1, Lv3, and Lv7 for video). The dashed blue lines represent isoquality lines in the absence of cross-modal interactions.

5.2 Cross-modal interactions for interactive testing

A similar analysis was performed for the interactive experimental context. In Figure 6, subjective scores for audio and video quality are represented for two levels of video quality and three levels of audio quality (La1, La4, and La7 for audio and Lv1 and Lv7 for video). An ANOVA computed on the video quality scores revealed only one significant difference between the extreme levels of audio quality (La1 and La7) and for high video quality (Lv7) with a MOS difference of 0.45. The type of conversational

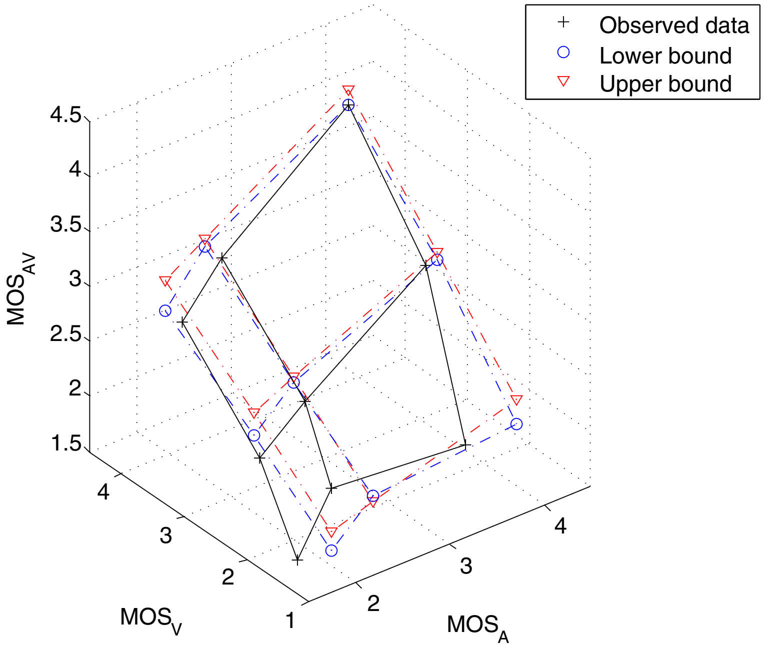

Figure 5 Impact of cross-modal interactions on the audiovisual integration for the passive experimental context (Exp. 1). Scatter plot of the subjective audiovisual scores obtained for different combinations of audio and video quality levels. The dashed blue and red lines represent the lower and upper bounds, respectively.

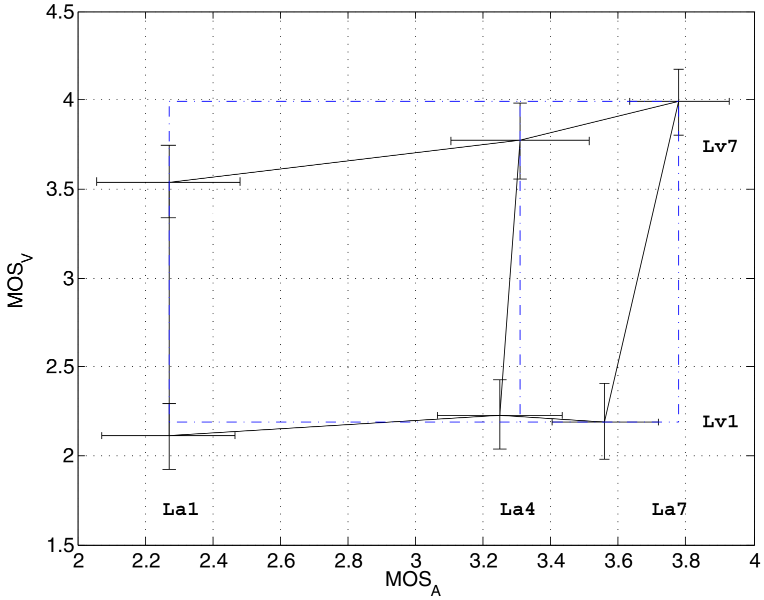

Figure 6 Cross-modal interactions for the interactive experimental context (Exp. 2). Scatter plot of the subjective audio and video scores with their 95% confidence intervals obtained for different combinations of audio and video levels. Vertical and horizontal lines connect data points obtained for similar audio and video quality levels, respectively (La1, La4, and La7 for audio and Lv1 and Lv7 for video). The dashed blue lines represent isoquality lines in the absence of cross-modal interactions.

scenario was added as an independent variable to the computation of the ANOVA to determine if the scenario influenced the result found for the main effect. A similar difference was detected for the SCT and AVSCT scenarios (MOS difference of 0.49 for the SCT scenario and of 0.65 for the AVSCT scenario) but was absent for the BB scenario. In contrast, no statistical difference in the audio MOS was detected due to variation of the video quality level. A trend for increase was observed for the level of high audio quality (La7) but did not reach the threshold of statistical significance. A scenario-wise analysis led to the same result for the SCT and AVSCT scenarios, and the trend for increase was only observed for the BB scenario but again was not significant. A summary of the significant cross-modal interactions detected for both experimental contexts can be found in Table 3.

The observed quality scores along with their lower and upper bounds are illustrated in Figure 7. The obtained results are similar to the ones found for the passive experimental context. The Pearson correlation computed between both bounds (0.98) and the low RMSE value (0.15) confirm the bound proximity found for the passive context. The influence of the audio quality level on the video MOS for high video quality leads to a clear offset between both bounds. The value of this offset depends on the level of audio quality but remains below 0.5 MOS. In this case, it can be considered as a minor influence on the audiovisual integration. However, if cross-modal effects would increase due to the conversational scenario, one could expect a greater impact on the audiovisual integration: it would cause the video channel to appear more correlated with the overall quality.

Table 3 Recapitulation of the cross-modal interactions for both experimental contexts and for the conversational scenarios

Context	Passive	Interactive			
		SCT	AVSCT	BB	All
A → V	x	x	x	none	x
V → A	x	none	none	none	none

5.3 Discussion

In this section, it was demonstrated that cross-modal effects principally occur in the high-quality domain. First, differences in quality ratings for one modality were statistically significant only when compared for extreme levels of the other modality. The level of audio quality had an impact on the perceived visual quality for both passive and interactive experimental contexts with an absolute value neighboring 0.5 MOS. However, the level of video quality only had an impact on the audio quality for the passive experimental context. Second, cross-modal effects do depend on the conversational scenario: the impact of the audio quality level on the video MOS was detected for the SCT and AVSCT scenario but not for the BB scenario. This corroborates the hypothesis that if a scenario requires an extensive usage of a particular channel, like

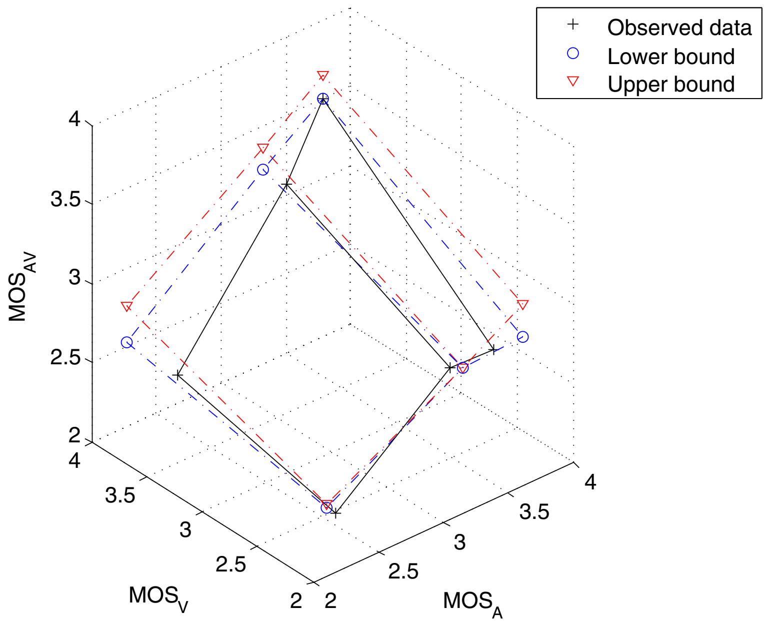

Figure 7 Impact of cross-modal interactions on the audiovisual integration for the interactive experimental context (Exp. 2). Scatter plot of the subjective audiovisual scores obtained for different combinations of audio and video quality levels. The dashed blue and red lines represent the lower and upper bounds, respectively.

the BB scenario with the video channel, the variation in quality of the other channel will not significantly affect its perception. Moreover, the fact that the level of video quality does not impact the perception of the audio quality in an interactive context (although the opposite case is true) suggests that the audio channel has a greater importance in a testing situation where test participants are not only placed in a judging position but are also asked to carry out a task for which they primarily rely on the audio channel to achieve. It could explain the trend in the audio ratings to be improved by the video quality level (even if not significantly) observed for the BB scenario.

For both experimental contexts, cross-modal effects turned out to have a very limited impact on the audiovisual integration. These results contrast with the work of Chateau [26] who found a strong impact of the video channel on the audio quality in an interactive context, but no cross-modal effects in a passive context. Two reasons can explain this difference of results: in his interactive experiment, the MOS range for the audio channel was very narrow, and the conversational scenario mainly involved the audio channel (Name Guessing Task).

6 Impact of the conversational scenario on the audiovisual modalities

In the last section, experimental evidence showed that cross-modal interactions could vary depending on the type of conversational scenario. This section will focus on measuring the actual influence of the conversational scenario on the evaluation of the perceived auditory and visual qualities. More precisely, each scenario involves different tasks that require using a channel more saliently than the other, both in terms of attention and information conveying. Interactants tend to adapt themselves and find strategies to convey the information in a visual manner if the auditory channel is impaired and vice versa [46]. Such adaptations could also lead interactants to shift their attention toward the channel predominantly used for carrying out the task and therefore constitute a significant cause of disparity in quality ratings between conversational scenarios.

6.1 Influence of the conversational scenario on auditory and visual qualities

Differences between conversational scenarios were observed when at least one channel was impaired. Quality ratings for a clean channel were similar, on average, across our conversational scenarios (see Figure 8 for quality levels La7 and Lv7). Starting with the impact of the scenario on the audio MOS, the SCT scenario was rated more negatively than the other scenarios in the case of strong audio degradations (La1 and La2) with an average difference of 0.58 MOS (computed over all video conditions). The difference in ratings between the SCT and the BB scenarios was statistically significant according to a *post hoc* test (Scheffé, $p < 0.05$). Conditions of no packet loss (La4 and La7) did not lead to any difference between the scenarios reflecting the fact that as long as the intelligibility of the

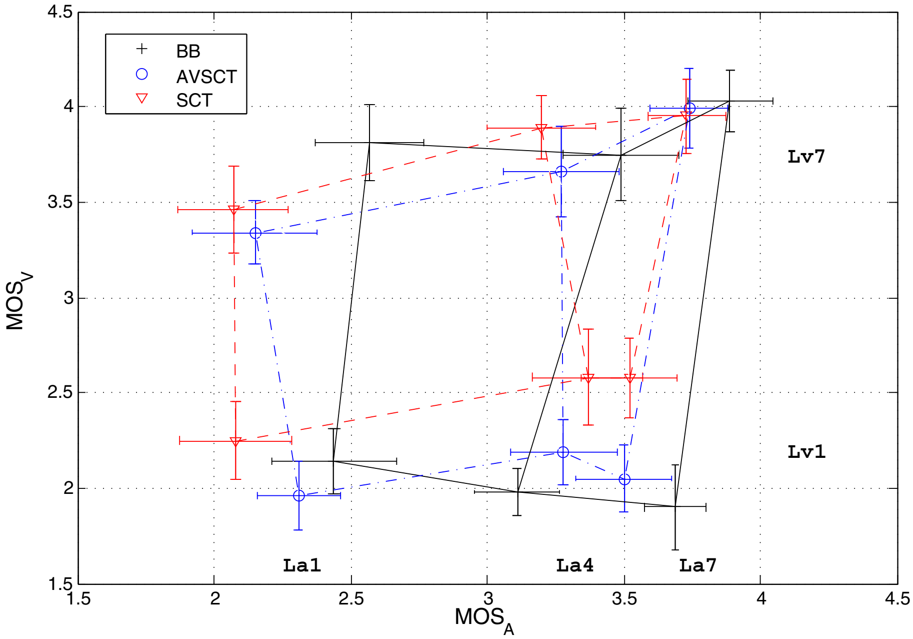

Figure 8 Cross-modal interactions for three conversational scenarios (Exp. 2). Scatter plot of the subjective audio and video scores with their 95% confidence intervals obtained for different combinations of audio and video levels. Vertical and horizontal lines connect data points obtained for similar audio and video quality levels, respectively (La1, La4, and La7 for audio and Lv1 and Lv7 for video).

speech signal was not affected, the type of scenario did not impact the perception of audio quality.

Concerning the impact of the scenario on the video MOS, the AVSCT and BB scenarios were judged more severely than the SCT scenario in the case of strong video degradations (Lv1) with a difference of 0.47 MOS (computed over all audio conditions). No difference was found for the other levels of video quality. It is interesting to notice that meaningful differences between scenarios only appeared for the combinations of a clear channel associated with a strongly impaired channel. The amplitude of the scenario's influence in terms of MOS value revolved around 0.5 MOS which is comparable to the influence of the cross-modal effects found in the last section. Finally, if a scenario did not require the usage of a channel (audio or video) for the task to be carried out, the quality ratings associated with this channel were more optimistic and inversely.

6.2 Impact of the conversational scenario on the audiovisual quality

Quality ratings for all conversational scenarios (introduced in Figure 8) are represented in the audiovisual space in Figure 9. No significant difference was observed for extreme levels of audiovisual quality, i.e., when both channels were left unimpaired (Lv7-La7) or were strongly impaired (Lv1-La1). However, when only one channel was strongly impaired and the other one was clean, differences could be observed: in the case of clean video combined with strongly impaired audio (Lv7-La1), the BB scenario received better audiovisual ratings as this latter drives the attention mostly on the video channel (MOS difference of 0.47 between the SCT and BB scenarios). Inversely, in the case of clean audio combined with strongly impaired video (Lv1-La7), the SCT scenario received better ratings with a MOS difference of 0.4. These differences were however not significant: one reason could be that only 16 ratings per conditions were gathered leading to a greater variance in quality judgements. Even if not statistically significant, the impact of the conversational scenario on the audiovisual quality will cause a higher apparent correlation of the audio dimension to the audiovisual one for the SCT scenario and a higher correlation for the video one for the BB scenario. The AVSCT scenario was equally affected by both modalities leading to a more balanced integration. The criterion differentiating the audiovisual integration between the scenarios is the correlation of the channel predominantly used for solving the task with the audiovisual quality as can be observed in Figure 9.

7 Application-oriented audiovisual integration

Predicting the audiovisual quality from the audio and video qualities was largely studied, and several quality-based models have been proposed. For such models, the perceived audiovisual quality (abr. MOS_{AV}) is expressed as a function of the audio and video qualities (respectively

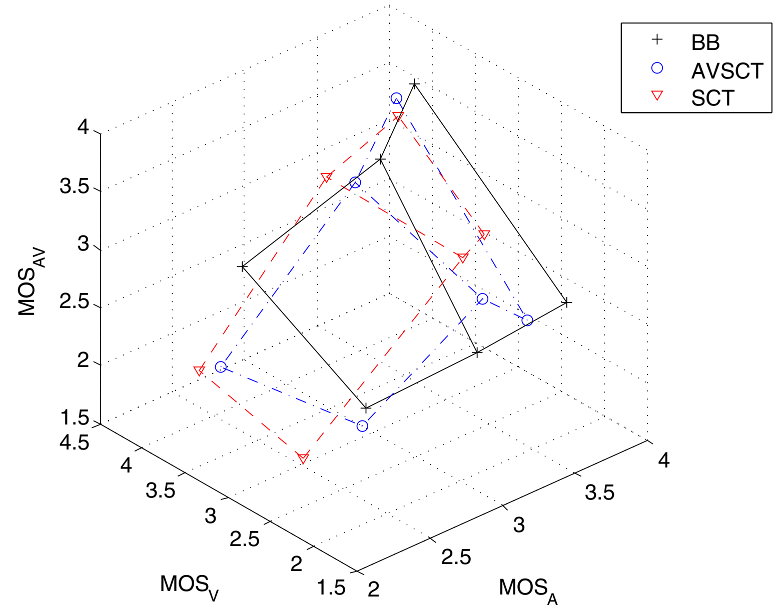

Figure 9 Impact of the conversational scenario on the audiovisual integration (Exp. 2).

MOS_A and MOS_V). A general model found in the literature reporting a good correlation between observed and predicted data is described by Equation 1. Some variants to this model only include a subset of the terms of Equation 1 depending on each term's contribution for explaining the variance of the experimental data. However, the direct impact of the visual and auditory modalities (or their interaction) on the audiovisual judgement depends on the type of application and on the audiovisual content. For television content, the video channel was found to be mostly predominant: the video quality strongly correlates with the audiovisual one [37]. Interactive services, like videotelephony, require a greater usage of the audio channel which will tend to emphasize the weight of the audio quality in the overall evaluation [29].

In this section, we will gather the audiovisual prediction models from the literature and from our experiments depending on the type of application and on the experimental context. First, we propose to define four categories to classify the different audiovisual services. Second, a single integration function is computed for each category. Third, we analyze the similarity between the individual integration functions and the models proposed category-wise. Two criteria were chosen to perform this analysis: (1) the Pearson correlation coefficient computed on the experimental data (usually reported in the literature) and the one computed between the predicted values of the individual models and of the proposed models for each category, and (2) an analytical distance between the individual integration functions and the categories' functions. Taken together, these two criteria provide a reliable indication of similarity between different integration functions.

7.1 Quality-based audiovisual metrics: literature review

Prior experiments on audiovisual integration are listed in Table 4. For each experiment, information concerning the laboratory and the year of publication are provided along with details concerning the experimental setup: the type of playout devices, the type of auditory and visual impairments, the employed testing methodology with the associated rating scale, and finally details concerning the audiovisual stimuli used in the experiments. The duration of the stimuli was generally comprised between 5 and 12 s except for experiment 2 where 25-s stimuli were used. For each experiment, a mathematical model for predicting the global audiovisual quality from the individual audio and video qualities could be derived. Several models based on Equation 1 were proposed for each experiment, and the coefficients are detailed in Table 5. As stated in [38], the multiplicative model usually reports a high correlation between observed and predicted data through a large variety of experimental conditions. However, this model can only achieve equal to lower performance in comparison to the model proposed in Equation 1 (see Table 5). This latter can explicitly reflect the weights of the individual correlations among the audio, video, and audiovisual qualities.

7.2 Experimental factors of influence

Factors related to the test design can be considered as having an impact on the experimental results: the

Table 4 Experiments on audiovisual integration from the literature

Category(index)	**Laboratory**	**Year**	**Playout devices**	**Degradation types**		**Scale, methodology**	**Audiovisual material**	**Reference**
				Video	**Audio**			
Television passive (1)								
1	Bellcore	1995	Television monitor and speakers	Random noise, simulated blurring, and blockiness	Temporally correlated noise, MNRU, and random noise	9-point, ACR	2 clips, 18 s Conversations in the street and in a library	[47]
2	KPN	1999	Television monitor and stereo loudspeakers	Spatial filtering of luminance	Temporal filtering (bandwidth limitation)	9-point, ACR	2 clips, 25 s Commercials	[33]
3	BT	2004	Television monitor and loudspeakers	Emulated blockiness	MNRU	5-point, SSQS	1 clip, 5 s Bicycle race with audio commentary	[29]
4	ICRFE	2005	Cellphone and handset	Coding	Coding	5-point, ACR	2 clips, 8 s Movie trailer, music clip	[36]
5	EPFL	2006	PC monitor and headphones	Coding	Coding	11-point, ACR	6 clips, 8 s Entertainment with speech and music	[35]
6	ITS	2010	HD monitor and loudspeakers	Coding	Coding	5-point, ACR	10 clips, 11 to 12 s Entertainment with speech and music	[38]
7	DT	2011	HD monitor and loudspeakers	Coding, packet loss, PLC	Coding, packet loss	11-point mapped to R-scale, ACR	5 clips, 16 s Entertainment with speech and music	[37]
Videotelephony passive (2)								
8	ITS	1998	PC monitor and speakers	Coding	Coding	5-point, ACR	6 clips, 5 to 9 s VTC (head-and-shoulders)	[48]
9	FT/CNET	1998	Television monitor and loudspeakers	Coding, frame rate, resolution	Coding	5-point mapped to 9-point, ACR	2 clips, 10 s VTC (head-and-shoulders)	[26]
10	BT	2004	Television monitor and loudspeakers	Emulated blockiness	MNRU	5-point, SSQS	1 clip, 5 s, VTC (head-and-shoulders)	[29]
11	ICRFE	2005	Cellphone and handset	Coding	Coding	5-point, ACR	1 clip, 8 s VTC (head-and-shoulders)	[36]
12	DT	2012	PC monitor and headphones	Coding, packet loss	Coding, packet loss	11-point mapped to 5-point, ACR	2 clips, 10 s VTC (head-and-shoulders)	Exp. 1

Table 4 Experiments on audiovisual integration from the literature *(Continued)*

Videotelephony interactive 1 (3)								
13	DT	2009	PC monitor and headphones	Coding, packet loss	Coding, packet loss	11-point mapped to 5-point, ACR	Scenario SCT, BB	[23]
14	DT	2012	PC monitor and headphones	Coding, packet loss	Coding, packet loss	11-point mapped to 5-point, ACR	Scenario SCT 3 min	Exp. 2
Videotelephony interactive 2 (4)								
15	NTT	2007	PC monitor and headphones	Coding, packet loss	Coding, packet loss	5-point, ACR	Scenario Name Guessing Task, 1 min	[49]
16	DT	2012	PC monitor and headphones	Coding, packet loss	Coding, packet loss	11-point mapped to 5-point, ACR	Scenario AVSCT, 3 min	Exp. 2
17	DT	2012	PC monitor and headphones	Coding, packet loss	Coding, packet loss	11-point mapped to 5-point, ACR	Scenario BB, 3 min	Exp. 2

Table 5 Overview of the audiovisual integration coefficients

	α	β	γ	ζ	ρ_{exp}	ρ_{max}/ρ_{min}	D_{min}/D_{max}
Television passive (1)							
1	0	0	0.114	1.912	0.99	0.99 (1) / 0.91 (3)	1.92 (1) / 4.56 (3)
2	0.007	0.24	0.088	1.12	0.98	0.99 (1) / 0.87 (3)	1.78 (1) / 5.70 (3)
	0	0	0.11	1.45	0.97	-	-
3	0	0.25	0.15	0.95	0.82	0.99 (1) / 0.87 (3)	3.60 (1) / 6.95 (3)
4	0.5691	0.5064	0.1697	−0.9222	0.91	0.99 (1) / 0.93 (3)	4.32 (1) / 7.06 (3)
	0	0	0.2329	0.9135	0.84	-	-
5	0.456	0.770	0	−1.51	0.94	0.98 (1) / 0.88 (3)	1.83 (1) / 4.89 (3)
	0	0	0.103	1.98	0.9	-	-
6	−0.0525	0.0274	0.1969	0.9845	0.96	0.99 (1) / 0.91 (3)	2.57 (1) / 6.01 (3)
	0	0	0.1919	0.9616	0.96	-	-
	0.6304	0.6807	0	−1.2757	0.94	-	-
7	0	0.13	0.006	28.49	0.94	0.99 (1) / 0.88 (3)	1.32 (1) / 4.56 (3)
	0	0	0.006	30.99	0.91	-	-
Model (1)	−0.001	0.101	0.158	1.101	0.98	-	-
Videotelephony passive (2)							
8	−0.0058	0.654	0.042	0.517	0.98	0.99 (2) / 0.92 (3)	2.02 (2) / 6.36 (3)
	0.217	0.888	0	−0.677	0.98	-	-
	0	0	0.121	1.514	0.93	-	-
9	0.35	0.57	0	−0.13	0.96	0.99 (2) / 0.91 (3)	1.28 (2) / 6.40 (3)
	0	0	0.10	1.76	0.96	-	-
10	0	0	0.17	1.15	0.85	0.99 (2) / 0.93 (3)	1.30 (2) / 5.07 (3)
11	0.2144	0.0124	0.1184	−0.6313	0.90	0.99 (2) / 0.92 (3)	3.93 (2) / 5.08 (3)
	0	0	0.9987	0.1536	0.89	-	-
12	0.114	0.242	0.103	0.887	0.93	0.99 (1) / 0.92 (3)	1.53 (1) / 4.02 (3)
Model (2)	0.020	0.046	0.141	1.162	0.98	-	-
Videotelephony interactive 1 (3)							
13	0.457	0.355	0.052	0.08	0.98	0.98 (3) / 0.82 (1)	2.09 (3) / 5.87 (1)
14	0.684	0.332	−0.034	0.209	0.99	0.99 (3) / 0.96 (2)	2.44 (3) / 6.10 (2)
Model (3)	0.584	0.357	0.0013	0.127	0.98	-	-
Videotelephony interactive 2 (4)							
15	−0.144	0.186	0.154	1.17	0.97	0.97 (4) / 0.93 (3)	2.47 (4) / 4.05 (2)
16	−0.038	−0.175	0.172	1.96	0.98	0.98 (4) / 0.99 (2)	1.79 (4) / 4.43 (2)
17	0.237	0.193	0.06	1.06	0.98	0.97 (1) / 0.81 (3)	3.23 (1) / 6.46 (3)
Model (4)	0.030	0.079	0.123	1.374	0.96	-	-

The original Pearson correlation of the integration functions with the experimental data is reported in the column ρ_{exp}. The column ρ_{max}/ρ_{min} reports the highest and the lowest Pearson correlation coefficients between the individual and the category functions associated with the category's index. The column D_{min}/D_{max} reports the smallest and the largest distances between the individual and the category functions associated with the category's index.

number, type and length of the stimuli, the range and type of impairments, the experimental context, the testing methodology, the number of test participants, and the targeted application. Four of these factors will be left out for the choice of the categories: (1) the variety of stimuli - the analysis realized here is performed at a coarse level, and the internal differences within one class of audiovisual contents are not taken into account; (2) the type and range of impairments - recent studies mainly include the same type of impairments, i.e., encoding-related degradations and losses of information packet for transmitted signals. The range of impairments is related to the balance of the test design, (3) the testing methodology - several commonly used methodologies and rating scales for audio and video quality assessment (SSCQE, DSCQS, ACR5, ACR11) were compared in [50-52], and results showed similarities between the testing methodologies; (4) the number of test participants - this factor influences the precision of the experimental data. We selected the experimental context and the targeted application as distinctive criteria for our categories. Results from Section 4 suggest that the experimental context is an influencing factor as the decrease of the audiovisual quality due to increasing the level of auditory and visual impairments was smaller in an interactive context than in a passive one, especially toward low visual quality. In Section 6, we showed that the impact of the conversational scenario had to be accounted for, and a distinction could be made between the SCT and BB scenarios.

Moreover, the contribution of both modalities to the overall quality can be measured by computing the linear Pearson correlation between those quantities. Table 6 summarizes the correlation coefficients obtained for prior experiments and for the ones presented in this study. For the evaluation of television contents in a passive context, either the video quality dominates the overall quality or both modalities have the same impact. Similar results are found for the evaluation of videotelephony contents in a passive context. For the interactive context, however, the SCT scenario showed a higher correlation of audio with the audiovisual quality than the video. We observe that the correlation of the audio quality to the audiovisual one increases from 0.62 to 0.93 between the BB and the SCT scenarios.

As a result, we propose to group the integration functions into four broad categories reflecting different types of application: (1) television contents, including various audiovisual material with potentially complex visual scenes and soundtracks (e.g., music); (2) videotelephony contents assessed in a passive context, characterized by a typical 'head-and-shoulders' scene and a fixed background, usually exhibiting low motion and scene complexity and accompanied with speech only; (3) videotelephony contents assessed in an interactive context including conversational scenarios mainly using the audio channel; and (4) videotelephony contents assessed in an interactive context including conversational scenarios mainly involving the video channel.

Table 6 Contribution of the audio and video qualities to the global audiovisual quality

Experiment	AV content	A/AV	V/AV
[28]	Television	0.33	0.90
[29]	Television	0.44	0.68
[35]	Television	0.55	0.67
[38]	Television	0.68	0.66
[37]	Television	0.47	0.80
[48]	Videotelephony passive	0.41	0.97
[26]	Videotelephony passive	0.49	0.88
[29]	Videotelephony passive	0.61	0.55
Exp. 1	Videotelephony passive	0.66	0.77
Exp. 2	Videotelephony interactive (SCT)	0.93	0.75
[23]	Videotelephony interactive (SCT, BB)	0.77	0.63
Exp. 2	Videotelephony interactive (AVSCT)	0.78	0.89
Exp. 2	Videotelephony interactive (BB)	0.62	0.89

Linear Pearson correlation coefficient between the subjective audio and audiovisual qualities (A/AV) and between the video and audiovisual qualities (V/AV).

7.3 Computational procedure for application-oriented integration functions

In order to derive a single integration function for each category, we used integration models reported from prior experiments in the literature and the ones derived from the experiments presented in this study. Table 5 provides the values of the audiovisual integration coefficients computed from the experimental data. The computation of a single integration function for each category was realized following a three-step procedure:

Step 1. Generating data points using the models specified in Table 5.

Step 2. Mapping the data points onto the 5-point ACR scale: according to Table 4, the scales used in these experiments were the 5-point, 9-point, and 11-point ACR scales and for one model, the R-scale. The first three scales are described in ITU-T Rec. P.910, and it is possible to map the scores from one scale onto another using a linear transformation[a]. The mapping from the R-scale to the 5-point MOS scale is defined in ITU-T Rec. G.107 [53].

Step 3. Computing a multiple regression analysis on the mapped data points using the model defined in Equation 1.

The obtained model's coefficients were added to Table 5 at the bottom of each category and designated by 'Model (x),' x being the corresponding category's index (1 to 4).

7.4 Performance evaluation

The application-oriented models aim at replacing the individual models for each proposed category. The underlying assumption is that the integration functions obtained for similar experimental contexts and audiovisual contents (or conversational scenarios) should present the same characteristics in the audiovisual space. Hence, the resulting audiovisual integration surfaces belonging to the same category should be spatially close to each other and can thus be merged into a unique function. One efficient way to directly calculate the distance between two integration functions is to evaluate the volume enclosed between the resulting integration surfaces. In the audiovisual space, the term 'close' means that for a given couple of audio and video MOS values, the corresponding audiovisual MOS values predicted by two individual functions f_i and f_j should be similar such as $f_i(\mathrm{MOS_A}, \mathrm{MOS_V}) \approx f_j(\mathrm{MOS_A}, \mathrm{MOS_V})$. It implies that the graphical representations of these functions should have a spatial proximity. In order to compute the volume, as a measure of distance between two integration functions, the following cases must be considered:

1. First case: the surfaces do not intersect each other. The volume between two functions is given by

$$D_{ij} = \int_1^5 \int_1^5 f_i(x,y) - f_j(x,y)\,\mathrm{d}x\,\mathrm{d}y \tag{2}$$

$$= 16(3\alpha_{ij} + 3\beta_{ij} + 9\gamma_{ij} + \zeta_{ij}), \tag{3}$$

   with D_{ij} being the volume between the functions f_i and f_j, $c_{ij} = c_i - c_j$ with $c \in \{\alpha, \beta, \gamma, \zeta\}$, and x and y being the variables of the audio and video quality dimensions, respectively.

2. Second case: the surfaces do intersect each other. The volume between the two functions is then given by

$$D_{ij} = \iint f_i(x,y) - f_j(x,y)\,\mathrm{d}\,R_1 - \iint f_i(x,y) - f_j(x,y)\,\mathrm{d}\,R_2, \tag{4}$$

   with R_1 being the region where $f_i(x,y) \geq f_j(x,y)$ and R_2 the region where $f_j(x,y) \geq f_i(x,y)$.

In the second case, an analytical solution of the double integral depends on the values of the c_{ij} coefficients. The resulting integral can be approximated using a global adaptive quadrature procedure in accordance with the separating function between the regions R_1 and R_2.

In order to evaluate the similarity between the category and the individual functions, we computed the linear Pearson correlation coefficients as well as the above-defined distance[b] between them and summarized the results in Table 5. It can be observed that the functions belonging to a same category are close to their corresponding category functions both in terms of correlation (correlation coefficients above 0.97) and distance measure. There were only two exceptions: functions n°12 and n°17 belonging to the categories 'videotelephony passive' and 'videotelephony interactive 2,' respectively, were closer to the category 'television.' Figure 10 shows the goodness of fit between the predictions of the individual and application-oriented functions. The resulting Pearson correlation coefficients values above 0.96, summarized in Table 5, demonstrate that the application-orientated models can efficiently replace the individual models.

8 Conclusion

The goal of this article was threefold: first, investigating the influence of audiovisual interactivity on QoE evaluation by comparing the results obtained for an interactive context to the classical passive viewing-and-listening context; second, analyzing the impact of the conversational scenario on the audiovisual quality evaluation; and third, proposing unified integration functions for audiovisual services jointly based on the results of the present study and from the literature. The analysis of two subjective experiments revealed a systematic impact of the experimental context: the variation of the MOS ranges shrank in passing from a passive to an interactive experimental context, which is in line with the results of Gros et al. for the evaluation of speech quality [14]. Noticeable differences in quality ratings were mainly found in the low-quality domain: in that case, ratings emitted in a passive context were more severe than in an interactive context. Cross-modal interactions occurred for both experimental contexts but were restricted to the high-quality domain. The audio quality level had a positive influence on the subjective video quality only when comparing extreme levels of audio quality and toward high video quality (for both experimental contexts). In turn, the video quality level had an impact on the subjective audio quality toward high audio quality and for the passive context only. The observed cross-modal interactions were measured to be on average of 0.5 on a 5-point MOS scale, which is in line with two studies from the literature [26,28] but contrasts with the larger value found in [27] (above 2 MOS points). The absence of significant influence of the audio quality level on the perceived video quality for both experimental contexts in [26] can be explained by the reduced range of variation in audio quality level used in that experiment. Cross-modal interactions also varied with the conversational scenario: an

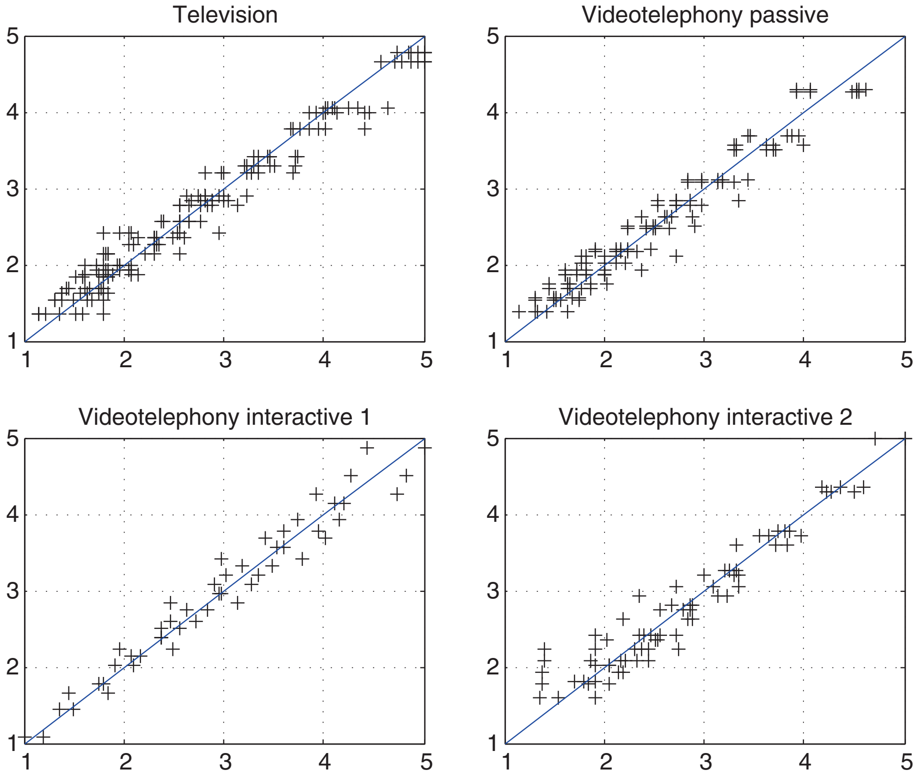

Figure 10 Correlation between the individual and application-specific models.

influence of the audio quality level on the video quality was detected for the SCT and AVSCT scenarios but not for the BB scenario. A comparison between three different interactive conversational scenarios with respect to their impact on the audio, video, and audiovisual qualities, respectively, showed that the conversational scenario directly influenced the quality evaluation. For the SCT scenario, the audio quality was rated more negatively in the case of strong audio degradations than for the other scenarios. The AVSCT and BB scenarios received poorer video quality ratings in case of strong video degradations. It is interesting to note that toward high video or audio quality, ratings for both modalities do not differ depending on the conversational scenario. The influence of the conversational scenario on the audiovisual quality was noticeable only when one modality was impaired: in the case of strong impairment of the audio channel, the BB scenario collected higher ratings; reciprocally, if only the video channel was strongly impaired, the SCT scenario obtained higher audiovisual ratings. Finally, an analysis of the models for audiovisual integration found in the literature combined with the results of the present experiments showed the relevance of deriving an integration function for each of the four proposed categories accounting for the experimental context, audiovisual content, and conversational scenario. Finally, in addition to propose several audiovisual integration functions for interactive video services, we derived an integration function for each type of application and showed that the resulting models can accurately replace the individual models from the literature without significantly affecting the prediction accuracy. These new models can thus be applied in a general context. The modeling approach as presented in this study focused on the 'satisfaction of a subject perceiving the signals regardless of what information is conveyed' [3]. A semantic analysis of the audiovisual content could help in determining which segments of the stimuli or of the conversation in an interactive context are more critical for the evaluation and thus should be attributed a greater weight in the modeling process.

Endnotes

[a] The linear mapping function is defined as follows: $Q_{\mathrm{new}} = (M_{\mathrm{new}} - m_{\mathrm{new}}) \cdot \frac{Q_{\mathrm{old}} - m_{\mathrm{old}}}{M_{\mathrm{old}} - m_{\mathrm{old}}} + m_{\mathrm{new}}$, where Q_{new} and Q_{old} represent the quality scores on the new and old scales, respectively; M_{new} and m_{new}, the maximal and minimal values on the new scale, respectively; and M_{old} and m_{old}, the maximal and minimal values on the old scale, respectively. After mapping, the quality scores exceeding the range $[M_{\mathrm{new}}, m_{\mathrm{new}}]$ were clipped.

[b] Let's note that this distance satisfies the four conditions defining a metric: 1. non-negativity: $D_{ij} \geq 0$,

2. identity: $D_{ij} = 0 \Leftrightarrow f_i = f_j$, 3. symmetry: $D_{ij} = D_{ji}$ and 4. triangle inequality: $D_{ik} \leq D_{ij} + D_{jk}$.

Competing interests

The authors declare that they have no competing interests.

Received: 3 March 2013 Accepted: 21 October 2013
Published: 9 November 2013

References

1. S Chikkerur, V Sundaram, M Reisslein, LJ Karam, Objective video quality assessment methods: a classification, review, and performance comparison. Broadcasting IEEE Trans. **57**(2), 165–182 (2011)
2. S Möller, WY Chan, N Cote, TH Falk, A Raake, M Wältermann, Speech quality estimation: models and trends. Signal Process. Mag. IEEE. **28**(6), 18–28 (2011)
3. J You, U Reiter, MM Hannuksela, M Gabbouj, A Perkis, Perceptual-based quality assessment for audio-visual services: a survey. Image Commun. **25**, 482–501 (2010)
4. MP Hollier, AN Rimell, DS Hands, RM Voelcker, Multi-modal perception. BT Technol. J. **17**, 35–46 (1999)
5. A Kohlrausch, S van de Par. Auditory-visual interaction: from fundamental research in cognitive psychology to (possible) applications. Proceedings of SPIE: human vision and electronic imaging IV , vol. 3466, San Jose, January 1999 (SPIE Bellingham), pp. 34–44
6. H McGurk, J MacDonald, Hearing lips and seeing voices. Nature. **264**, 746–748 (1976)
7. A Joly, N Montard, N Buttin. Audio-visual quality and interactions between television audio and video. Sixth International Symposium on Signal Processing and Its Applications, vol. 2 (Kuala Lumpur, August 2001)
8. N Kitawaki, K Itoh, Pure delay effects on speech quality in telecommunications. Selected Areas Commun. IEEE J. **9**(4), 586–593 (1991)
9. ITU-T Recommendation P920. Interactive test methods for audiovisual communications. Technical report. International Telecommunication Union, Geneva (2000)
10. F Hammer. Quality aspects of packet-based interactive speech communication. PhD thesis, University of Technology Graz, Graz (2006)
11. M Guéguin, R Le Bouquin-Jeannès, V Gautier-Turbin, G Faucon, V Barriac, On the evaluation of the conversational speech quality in telecommunications. EURASIP J. Adv. Signal Process. **2008**, 185248 (2008)
12. S Egger, R Schatz, S Scherer. It takes two to tango - assessing the impact of delay on conversational interactivity on perceived speech quality. 11th Annual Conference of the International Speech Communication Association (Interspeech), Makuhari, September 2010 (ISCA, Washington D.C, 2010)
13. K Yamagishi, T Hayashi, Analysis of psychological factors for quality assessment of interactive multimodal service. Hum. Vis. Electron. Imaging X. **5666**, 130–138 (2005)
14. L Gros, N Chateau, S Busson, Effects of context on the subjective assessment of time-varying speech quality: listening/conversation, laboratory/real environment. Acta Acustica U. Acustica. **90**(6), 1037–1051 (2004)
15. S Möller, *Assessment and Prediction of Speech Quality in Telecommunications*. (Kluwer, Boston, 2000)
16. U Reiter, J You. Estimating perceived audiovisual and multimedia quality—a survey. IEEE 14th International Symposium on Consumer Electronics (ISCE), Braunschweig, June 2010 (IEEE, Piscataway, 2010), pp. 1–6
17. WR Neuman, AN Crigler, WM Bove. Television sound and viewer perceptions, in *Audio Engineering Society Conference: 9th International Conference: Television Sound Today and Tomorrow,* Detroit, 1–2 Feb 1991 (Audio Engineering Society New York, 1991), pp. 101–104
18. AQ Summerfield, Use of visual information in phonetic perception. Phonetica. **36**, 314–331 (1979)
19. DA Slutsky, GH Recanzone, Temporal and spatial dependency of the ventriloquism effect. NeuroReport. **12**, 7–10 (2001)
20. RB Welch, DH Warren, Immediate perceptual response to intersensory discrepancy. Psychol. Bull. **88**(3), 638–667 (1980)
21. A Kohlrausch, S van de Par, *Audio-visual Interaction in the Context of Multi-media Applications*. (Springer, Berlin, 2005), pp. 109–138
22. B Julesz, IJ Hirsch, ed. by EE David, PB Denes. Visual and auditory perception - an essay of comparison, in *Human Communication: A Unified View* (McGraw-Hill New York, 1972), pp. 283–340
23. B Belmudez, S Möller, B Lewcio, A Raake, MA Mehmood. Audio and video channel impact on perceived Audio and video channel impact on perceived audio-visual quality in different interactive contexts. IEEE International Workshop on Multimedia Signal Processing (MMSP), Rio de Janeiro, October 2009 (IEEE, Piscataway, 2009)
24. S Möller, B Belmudez, MN Garcia, C Kühnel, A Raake, B Weiss. Audio-visual quality integration: comparison of human-human and human-machine interaction scenarios of different interactivity. Second International Workshop on Quality of Multimedia Experience (QoMEX), Trondheim, June 2010 (IEEE Piscataway, 2010)
25. D Kahneman, *Attention and Effort*. (Prentice Hall, Upper Saddle River, 1973)
26. ITU-T Contribution COM 12-61-E, Study of the influence of experimental context on the relationship between audio, video and audiovisual subjective qualities. Technical report. International Telecommunication Union, Geneva (1998)
27. AN Rimell, MP Hollier, RM Voelcker. The influence of cross-modal interaction on audio-visual speech quality perception, in *Audio Engineering Society Convention 105,* San Francisco, 26–29 Sept 1998 (Audio Engineering Society New York, 1998)
28. FE Beerends, The influence of video quality on perceived audio quality and vice versa. J. Audio Eng. Soc. **47**(5), 355–362 (1999)
29. DS Hands, A basic multimedia quality model. IEEE Trans. Multimedia. **6**(6), 806–816 (2004)
30. ANSI-Accredited Committee T1 Contribution, Report on an experimental combined audio/video subjective test method. Technical report T1A1.5/93-104, Bellcore, Red Bank (1993)
31. ANSI-Accredited Committee T1 Contribution, Report on extension of combined audio/video quality model. Technical report T1A1.5/94-141, Bellcore, Red Bank (1993)
32. ITU-T Contribution COM 12-64-E, Results of an audiovisual desktop video teleconferencing subjective experiment. Technical report. International Telecommunication Union, Geneva (1998)
33. ITU-T Contribution COM 12-19-E, Relations between audio, video and audiovisual quality. Technical report. International Telecommunication Union, Geneva (1997)
34. ITU-T Recommendation P911, Subjective audiovisual quality assessment methods for multimedia applications. Technical report. International Telecommunication Union, Geneva (1998)
35. S Winkler, C Faller, Perceived audiovisual quality of low-bitrate multimedia content. Multimedia. IEEE Transactions. **8**(5), 973–980 (2006)
36. M Ries, R Puglia, T Tebaldi, O Nemethova, M Rupp, Audiovisual quality estimation for mobile streaming services. 2nd International Symposium on Wireless Communication Systems, Siena, September 2005, 173–177 (2005)
37. MN Garcia, R Schleicher, A Raake, Impairment-factor-based audiovisual quality model for IPTV: influence of video resolution, degradation type, and content type. EURASIP J. Image Video Process. **2011**, 629284 (2011)
38. MH Pinson, W Ingram, A Webster, Audiovisual quality components. Signal Process. Mag. IEEE. **28**(6), 60–67 (2011)
39. ITU-T Recommendation P911, Subjective audiovisual quality assessment methods for multimedia applications. Technical report. International Telecommunication Union, Geneva (1998)
40. B Lewcio, S Möller. A testbed for QoE-based multimedia streaming optimization in heterogeneous wireless networks. IEEE 5th International Conference on Signal Processing and Communication Systems (ICSPCS), Honolulu, December 2011 (IEEE Piscataway, 2011), pp. 1–7
41. Website of PJSIP, an Open Source SIP stack and Media Stack for Presence Instant Messaging and Multimedia Communication. http://pjsip.org/. Accessed 20 Jan 2009
42. Website of ffmpeg. http://www.ffmpeg.org/. Accessed 7 June 2008
43. Netem – The Linux Foundation. http://www.linuxfoundation.org/collaborate/workgroups/networking/netem. Accessed 20 Jan. 2009
44. ITU-T Recommendation G1071, Wideband e-model. Technical report. International Telecommunication Union, Geneva (2011)
45. ITU-T Contribution COM 12-C271, Description of conversation scenarios for interactive audiovisual communication. Technical report. International Telecommunication Union, Geneva (2011)

46. S Egger, P Reichl, M Ries. Quality-of-experience beyond MOS: experiences with a holistic user test methodology for interactive video services, in *21st ITC Specialist Seminar on Multimedia Applications - Traffic, Performance and QoE* (Miyazaki, 2–3 March 2010), pp. 13–18
47. ANSI-Accredited Committee T1 Contribution, Combined A/V model with multiple audio and video impairments. Technical Report T1A1.5/94-124, Bellcore, Red Bank (1995)
48. C Jones, DJ Atkinson. Development of opinion-based audiovisual quality models for desktop video-teleconferencing. Sixth International Workshop on Quality of Service (IWQoS), Napa, May 1998 (IEEE Piscataway, 1998), pp. 196–203
49. T Hayashi, K Yamagishi, T Tominaga, A Takahashi. Multimedia quality integration function for videophone services. IEEE Global Telecommunications Conference. GLOBECOM '07, Washington, DC, November 2007 (IEEE, Piscataway, 2007), pp. 2735–2739
50. M Pinson, S Wolf. Comparing subjective video quality testing methodologies. SPIE Video Communications and Image Processing Conference, VCIP 2003, Lugano (SPIE, Bellingham, 2003)
51. T Tominaga, T Hayashi, J Okamoto, A Takahashi. Performance comparisons of subjective quality assessment methods for mobile video. Second International Workshop on Quality of Multimedia Experience (QoMEX), Trondheim, June 2010 (IEEE Piscataway, 2010), pp. 82–87
52. S Zielinski, P Brooks, F Rumsey. On the use of graphic scales in modern listening tests, in *Audio Engineering Society Convention 123* (New York, 5–8 October 2007)
53. ITU-T Recommendation G107, The E-model, a computational model for use in transmission planning. Technical report. International Telecommunication Union, Geneva (2005)

doi:10.1186/1687-4722-2013-24
Cite this article as: Belmudez and Möller: **Audiovisual quality integration for interactive communications.** *EURASIP Journal on Audio, Speech, and Music Processing* 2013 **2013**:24.