Chapter 2
Interactive Model Framework

Abstract Evaluating user perception of audiovisual interactive services like videotelephony in a reliable fashion calls for a well controlled testing environment and experimental test bed. The present chapter introduces the main aspects of the employed experimental method for studying the perception of audiovisual quality for videotelephony. In this work, a dedicated test bed was deployed that is composed of a controlled laboratory environment, a network infrastructure, a videotelephony client and a control unit. Audiovisual material specific to videotelephony ("head-and-shoulders") was produced following specific conversational scenarios adapted to the evaluation of the interactive quality. This experimental setup was designed to facilitate the investigation of user experience in an interactive experimental context.

2.1 Modeling Framework

This book focuses on improving several aspects of the parameter-based model ITU-T Rec. G.1070. This model is used for network planning (see Sect. 1.5.1), which corresponds to the case where the service under study is not set up yet. The network planner must decide on the characteristics of the elements being part of the end-to-end transmission chain. As mentioned in Sect. 1.3, the elements impacting quality the most are the audio and video codecs (and their associated profiles), the operating bit rate, the parameters controlling the signal acquisition and the network packet loss rate. The model takes these parameters as input and provides three quality scores: the audio quality, the video quality and the audiovisual quality. These quality scores predict the opinion that would have been emitted by a user of the service for short audiovisual samples (app. 10 s) assuming a constant temporal quality profile.

In this chapter, an experimental test bed for assessing interactive audiovisual quality of videotelephony will be presented. On one hand, the test bed allows to process video sequences by introducing realistic impairments and on the other hand it can be used in real-time to implement a video call between two VVoiP clients. Figure 2.1 illustrates the three main steps of the quality assessment process that will be investigated, namely, the evaluation of the single modalities, the audiovisual integration, and finally the temporal pooling.

B. Belmudez, *Audiovisual Quality Assessment and Prediction for Videotelephony*,
T-Labs Series in Telecommunication Services, DOI 10.1007/978-3-319-14166-4_2

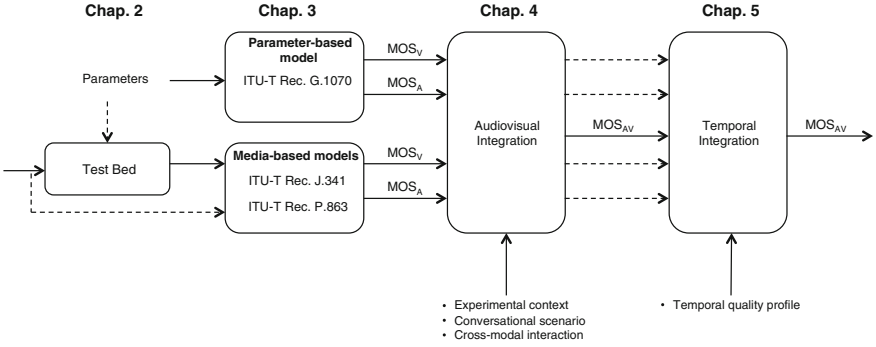

Fig. 2.1 Block diagram for video call quality assessment

Table 2.1 Overview of subjective quality scores databases

Database	AV content	Presentation mode	Context	Chapters	Publication
VO-VTp-1	Videotelephony	Viewing-only	Passive	3	[17, 63, 69]
VO-VTp-2	Videotelephony	Viewing-only	Passive	3	[63]
VO-MoTVp-1	Mobile TV	Viewing-only	Passive	3	–
VO-MoTVp-2	Mobile TV	Viewing-only	Passive	3	[18, 61]
VO-MoTVp-3	Mobile TV	Viewing-only	Passive	3	[61]
AV-VTp-1	Videotelephony	Viewing and listening	Passive	4, 5	[16, 19, 64]
AV-VTp-2	Videotelephony	Viewing and listening	Passive	4	–
AV-VTsi-1	Videotelephony	Viewing and listening	Semi-interactive	5	[16]
AV-VTsi-2	Videotelephony	Viewing and listening	Semi-interactive	5	[16, 114]
AV-MoTVp-1	Mobile TV	Viewing and listening	Passive	4	–
AV-MoTVp-2	Mobile TV	Viewing and listening	Passive	4	–
AV-VTi-1	Videotelephony	Viewing and listening	Interactive	4	[20]
AV-VTi-2	Videotelephony	Viewing and listening	Interactive	4	[66]
AV-VTi-3	Videotelephony	Viewing and listening	Interactive	4	[19, 64]

In order to investigate each block of the video call quality assessment process, a set of subjective experiments was conducted with the developed test bed, comprising 5 video-only experiments and 9 audiovisual experiments. The results of these experiments constitute the databases, presented in Table 2.1, which are used in the next chapters for both modeling and performance evaluation purposes. The databases are classified according to the mode of presentation (viewing-only or viewing and listening), the experimental context (passive, semi-interactive or interactive) and the audiovisual content (videotelephony or Mobile TV). Publications related to these databases are additionally listed. A complete description of the test plan as well as the collected subjective ratings for each experiment can be found in the Appendix C.

Table 2.2 Quality elements of the transmission chain controllable through the videotelephony client

Acquisition	Encoding	Packetization	Transmission	Buffering	Decoding	Playout
Video						
Resolution	Codec (profile/level)	Frame slicing	Network type	Buffer size	PLC	Image size
Format (chroma sampling)	Bit rate	Packet size	Delay			
	Quantization		Jitter			
Frame rate	GoP size		Packet loss rate			
Audio						
Sampling rate	Codec	Packet size	id. Video	id. Video	id. Video	Loudness
	Bit rate					

In Chap. 3, impaired audio and video sequences are produced to reflect different use cases of videotelephony transmissions by varying parameters of the application and network layers. An exhaustive overview of the parameters is given in Table 2.2. The main investigated parameters are the audio and video codecs, the video encoding resolution, the video display size, the video operating bit rate, the video frame rate and the network packet loss rate. The quality of these sequences is retrospectively assessed by test subjects who provided a MOS score. The database VO-VTp-1 is used to derive the coefficients of the G.1070 video quality function for several video codecs. The values of the coefficients are validated against two databases, VO-VTp-2 and VO-MoTVp-1. i.e. for videotelephony contents but also for Mobile TV contents exhibiting a larger variety of spatial and temporal complexity. Media-based models like J.341 for video and P.863 for audio (see description in Sect. 1.5.2) are used on videotelephony databases VO-VTp-1 and VO-VTp-2 to compare performances between parametric and media-based models. Two additional viewing-only experiments are conducted (VO-MoTVp-2 and VO-MoTVp-3) for investigating the impact of encoding resolution and display size on the subjective video quality and proposing an extension to the G.1070 video quality estimation function. The database VO-MoTVp-2 represents the training test for the G.1070 extension modeling as VO-MoTVp-3 serves the sole purpose of verifying the validity of VO-MoTVp-2. The modeling approach is cross-validated by dividing the training database in two non-overlapping datasets.

In Chap. 4, models of audiovisual quality are derived based on subjectives quality scores. They correspond to several use cases: on one hand, a passive situation of assessment with either videotelephony content (AV-VTp-1, AV-VTp-2) or Mobile TV content, (AV-MoTVp-1 and AV-MoTVp-2) and on the other hand, an interactive situation of assessment with different conversational scenarios (AV-VTi-1, AV-VTi-2 and AV-VTi-3). The performance of these models is then assessed on audio and video

scores provided by both types of predictive models: parameter-based for passive and interactive experimental contexts and media-based for the passive context.

In Chap. 5, temporal aspects of audiovisual integration are investigated. The quality of conversational quality is predicted based on momentary quality ratings. These ratings are either subjective or predicted by media-based models and are taken from database AV-VTp-1. Two supplementary databases are used for obtaining quality measurements of entire video calls (AV-VTsi-1 and AV-VTsi-2). AV-VTsi-1 is built using stimuli produced for AV-VTp-1 and is used for the optimization of temporal quality models. AV-VTsi-2 is an independent database containing similar audiovisual contents as AV-VTsi-1 with identical temporal length but different types of degradations, and is used for validating the proposed modeling.

2.2 Experimental Setup

Most publicly available videotelephony clients provide a limited user control over the parameters controlling the characteristics of the audio and video channels. Those parameters were referred to as the quality elements in Chap. 1. When investigating the quality impact of specific parameters, it is necessary to independently and accurately attune their control range. Therefore, a modular videotelephony client was especially developed for the experimental purposes of this work enabling a full control of the audio and video streams independently.

2.2.1 Videotelephony Software

The technical parameters for both the audio and video channels that are made available to the experimenter in the videotelephony software are summarized in Table 2.2. This software is based on a VoIP client project called PJPROJECT 0.8.3 [3]. It uses the SIP [142] and RTP/RTCP [145] protocols to manage multimedia sessions and transmit data in real-time, respectively. This open-source framework was chosen for its modularity that allowed to develop supplementary modules that could be used for research purposes. In the initial project, only an audio media flow was available for transmitting encoded speech (VoIP). A video channel was developed as a second independent media flow and integrated as part of this work. Video encoding and decoding was based on open libraries such as libavformat, libswscale and, libavcodec [2].

As can be seen in Fig. 2.2, the media flow of the software is quite similar to the general diagram provided in Fig. 1.5. The media streams for audio and video have a similar structure: they are composed of a signal bridge (Audio or Video bridge), that interconnects multiple sources like raw signals from peripheral devices (camera, microphone), file I/O, and the media streams. A media stream is created for each multimedia session and consists of a codec (encoder/decoder), a jitter buffer, and

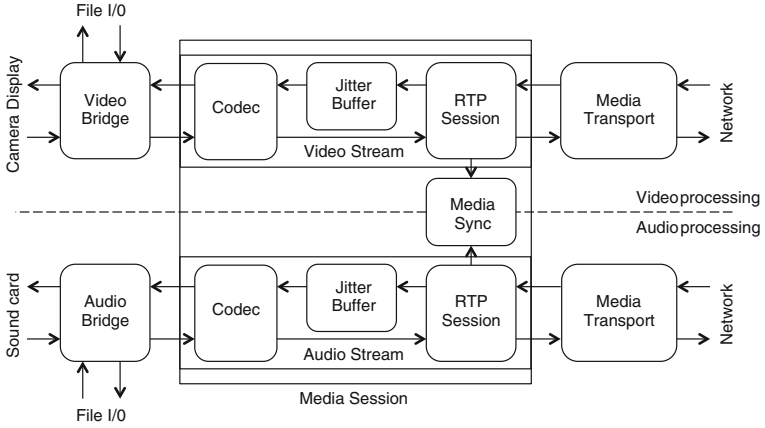

Fig. 2.2 Block diagram of the videotelephony software media flow (adapted from [116])

a RTP session module. A media transport unit controls the network sockets of the RTP/RTCP protocol. The media sync module is used to synchronize the audio and video media streams thanks to the timing information of the RTP timestamps. On the sender side, audio and video encoders are used to encode the raw audio and video signals to produce a bitstream called Elementary Stream (ES) for each signal. This stream is then packetized following RFC standards[1] and sent over the network by the media transport unit. It is important to mention that both audio and video streams are sent separately as for some multimedia applications they are multiplexed. On the receiver side, the packets are decapsulated upon arrival and their payload transferred to a jitter buffer where they are reordered according to the RTP sequence number. All payloads belonging to a single frame are then passed to the decoder that uncompresses the frame and passes it to the signal bridge for playout.

The software controls the acquisition parameters of the audio and video signals. The initialization of the peripheral devices allows to set the audio clock rate (from 8 to 48 KHz), the video color space (for more information on the available video color space, see [116]), the video resolution (up to *VGA*), as well as the video frame rate (from 5 to 30 fps). The signals can be dynamically routed to different interfaces: the audio signal to the sound card and the video signal sent to different Graphical User Interfaces (GUIs). Moreover the uncompressed signal (after decoding) can be written to a file with an AVI format[2] in order to grasp all degradations that affected the audiovisual stream until presentation to the test subjects. This allows to capture the signal as it would be viewed/heard by the user, i.e. including degradations like video freezing (in case of an empty jitter buffer) or rescaling artifacts. As mentioned earlier,

[1] The following RFC standards were used to packetize video streams: RFC 2250 for MPEG-2 [52], RFC 6416 for MPEG-4 [143] and RFC 6184 for H.264 [156].

[2] The AVI container was chosen as it is adapted to store uncompressed audio and video streams along with the associated metadata.

the client can principally be used in two modes: “off-line” mode for the production of degraded audiovisual sequences for usage in non-interactive subjective tests, and “on-line” mode for interactive test sessions.

At the encoding stage, a large variety of speech and video codecs can be employed to compress the media streams. Standardized speech codecs for narrowband (0.3–3.4 kHz), wideband (0.05–7 kHz), superwideband (0.05–14 kHz) and full band (0.02–20 kHz) are available. The FFmpeg library provides open access to numerous video codecs, including MPEG-2, H263+, MPEG-4 Part 2 and H.264. All codec features made available by the codecs are adjustable. Among these, are the codec profile, the operating bit rate, the quantization parameter, and the Group of Pictures size.

At the packetization and transmission stages, several parameters are accessible: the slicing parameter of the video frames, the packet size, the type of network (wired or wireless connections), the delay between packet emission and reception, and random packet loss rate. The network related parameters are controlled by the test application through a network emulation tool (Netem [1]), acting as a filter of the outgoing packets on the network interface. A full description of the software capabilities can be found in [114, 116].

2.2.2 Test Controller

The control of the videotelephony client is handled by a separate module called “Test Controller”. This control unit enables the automatic processing of predefined operations. For passive testing, the client is used to produce a corpus of stimuli (processed audiovisual sequences) according to specific test conditions. In that case, the test controller reads an XML file containing the description of all test conditions under study and collect the input parameters for each test condition. The audiovisual files are then processed according to the instructions provided by the Test Controller. It starts by initializing both instances of the clients, i.e. audio and video acquisition parameters, media codecs, network parameters (type of network, packet loss rate). It then automatically reads the uncompressed video files and processes them through the test bed. The files are recorded at the end of processing chain, i.e. at the playout stage. The Test Controller will read and execute every condition contained in a descriptive XML file.

Figure. 2.3 gives an example of such a descriptive XML file. The audio and video files to be read are specified first, followed by the type of network (LAN), the video and audio codecs with their associated bit rate (e.g. video: H.264@512 kbps, audio: G.722@64 kbps) and finally the network packet loss values for each channel (video 3 %, audio: 20 %).

For interactive testing, the client is jointly controlled with a GUI specifically designed for video call experiments. The GUI was developed as a shared-object library and matched the API of the test client in order to be dynamically loaded. It guides the participants through the test session, allowing them to be autonomous, i.e., controlling when the conversation should start and end. Rating scales for the

```
<test_condition>
    <audio_file>audio_input.wav</audio_file>
    <video_file>video_input.yuv</video_file>
    <output_file>output.avi</output_file>
    <network>
        <value>LAN</value>
    </network>
    <video_codec>
        <value>H264_512</value>
    </video_codec>
    <audio_codec>
        <value>G722_64</value>
    </audio_codec>
    <audio_loss>
        <value>20</value>
    </audio_loss>
    <video_loss>
        <value>3</value>
    </video_loss>
</test_condition>
```

Fig. 2.3 Example of an XML file defining experimental test conditions

evaluation phase were automatically displayed after each conversation. The Test Controller updates the test bed settings before each new conversation so that test subjects experience the different transmission characteristics defined by the test conditions. Finally, it saves the quality ratings in an XML file.

2.2.3 Rating Scales

For all experiments listed in Table 2.1, an Absolute Category Rating method was used for the subjective assessing of quality (see Sect. 1.4.1). Despite its drawbacks, this method is widely used for assessing quality in the telecommunication domain. Moreover, the ACR method was used for developing the G.1070 model. Experimental methodology comparisons proved that this method yields a good repeatability and is efficient as each stimulus should only be seen/viewed once. Besides, an absolute rating method is preferable as it represents a real situation of assessment were users have to emit an "absolute" opinion.

The continuous 11-point scale [93] (see Fig. 1.7) was used. This scale produces similar results to the 5-point MOS scale but attenuates certain drawbacks of the 5-point scale. For instance, the separate extremities above the numeric label 9 and below the label 1 tend to reduce the saturation effect (see Sect. 1.4.3) and the use of numbers along with the labels consolidates the "interval" characteristic of the scale [119]. An example of the implementation of the 11-point continuous rating scale in the experimental GUI for interactive testing is shown in Fig. 2.4. The quality labels for individual rating categories were given in German. The collected ratings were linearly mapped to the 5-point ACR category scale.

Fig. 2.4 11-point continuous scales used for interactive audiovisual quality assessment

2.2.4 Experimental Environment

The listening and viewing conditions were compliant with ITU-T Recommendations P.800 for listening tests [86], P.910 for viewing tests [93], and P.911 for audiovisual tests [94]. Careful attention had to be given to the following factors: the lightening of the room, the luminance and gamma values of the screen and the playout speech level. D65 chromaticity lights (temperature of 6,504 K) were used to realize a daylight illumination of the testing room. The walls were uniformly gray to avoid disturbing color perception on the screen. The background of the screen was homogeneously lit to respect a value of 20 cd/m^2. The gamma value of the display was set to 2.2, through the video card driver. The speech level of the headphones was calibrated to reach a value of approximately 80 dB(A) [95]. The rooms were sound insolated to block noise from the outside environment that would impair the listening conditions (ambient noise level below 30 dB(A)). 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).

2.2.5 Test Subjects

The recruited participants for subjective tests were balanced in gender and aged between 18 and 40. They were not concerned with multimedia quality as part of their work, and therefore were not experienced assessors. Prior to each test session, the observers were screened for normal visual acuity or corrected-to-normal acuity and for normal color vision. Moreover, they received monetary compensation for their participation. Subjects were given instructions on how to properly use the rating scales during a test session. However, it can always occur that some subjects concentrate their judgements toward a specific part of the scale, which results in a skewed distribution of the subjective scores.

As a result, the subjective scores of the participants have to be screened to detect any abnormal behavior. First, a cross-correlation between the scores of the different test participants gives an indication on the similarity of the score distributions. Generally, the cross-correlation coefficient should be above 0.7 for a sufficient confidence

in the participants rating's behavior. Another useful indicator is the Cronbach's alpha coefficient that can be calculated on the test participants. It is a reliability measure that expresses the internal consistency of the subjects' group. When these two indicators coincide, i.e. poor correlation of one subject's ratings with the rest of the group (Pearson correlation below 0.7) and a Cronbach's alpha coefficient that increases if the subject's ratings are omitted, then the subject should be removed from the database.

2.3 Quality Evaluation and Level of Interactivity

Figure 2.5 depicts the different types of audiovisual stimuli and conversational scenarios developed in this book for quality assessment purposes depending on the modality under test and the level of interactivity. Several types of stimuli, including interactive scenarios, were developed to allow a comparison between experimental results from different subjective tests. First, an adaptation for the audiovisual case was made of the simulated conversational structures developed by Weiss et al. [157] for the assessment of semi-interactive speech dialogs. An audiovisual simulated conversation is composed of several sequences containing a speaker's head and torso, uttering a sentence about a specific topic (e.g. renting a car). The speaker simulates one dialog partner of a conversation, as the other partner is the actual test subject who is asked to answer a question after each sequence in order to get involved in the actual content of the sequence, like in a real conversation. The simulated conversations are by design separable into semantically independent units of about 10 s that can serve as short samples for passive listening and viewing testing. Moreover, the scenarios used for the simulated conversations could be adapted into actual interactive conversational scenarios by inserting a structure alternating questions and answers between

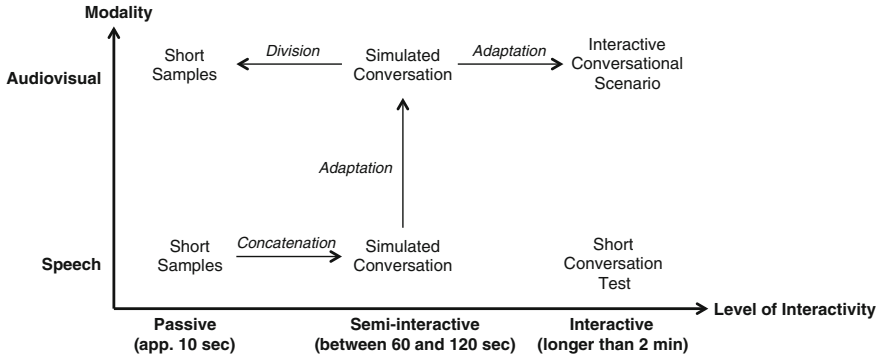

Fig. 2.5 Development scheme of the audiovisual stimuli and conversational scenarios depending on the level of interactivity. The direction of the arrows indicates the order of development to respect a semantical homogeneity across different levels of interactivity

two conversing partners. The advantage of proceeding that way was that the topics used in the simulated conversations could be kept along with the associated material (items, calendar, etc.). Therefore, the experienced audiovisual content generated during an interactive conversation was expected to be close to the one of a simulated conversation in terms of content, i.e. spatial and temporal complexity, and by extension close to the short samples as well.

2.4 Audiovisual Stimuli

2.4.1 Short Samples

Two types of video sequences were used in the passive experiments that reflected different types of applications, namely Mobile TV (MoTV) and videotelephony (VT). Four sequences for Mobile TV were taken from HDTV applications and resized from *H D* (1920 × 1080) to *VGA* (640 × 480) format and shortened from 16 to 10 s. They were representative of the service usage by their diversity in content including, a movie trailer with music, an interview with speech (similar to news content), a music clip and a soccer game with crowd noise.

The sequences for videotelephony were produced in accordance with the topics of the simulated conversations like the storytelling of a birthday party, a car rental, making a doctor appointment and purchasing a kitchen. In order to produce stimuli which are meaningful for an audiovisual conversation, the simulated telephone conversations used in [157] were modified by adding a video channel with visual cues (i.e. showing objects to the camera, pointing dates on a wall calendar, body gestures), so that the test participants had to pay attention to the video channel. For each of these scenarios, ten short samples were produced by a different German speaker (2 males and 2 females) with two different scene backgrounds. Each sentence part of a simulated conversation (short sample) was recorded in raw format (uncompressed planar YUV 4:2:0) with a *VGA* 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. The audiovisual content of the video sequences is detailed in Table 2.3.

As indicators of the spatial and temporal complexity of the scenes, the spatial perceptual information (SI) and temporal perceptual information (TI) defined in ITU-T Rec. P.910 were calculated following Eqs. (1.1) and (1.2) respectively. Figure 2.6 displays the SI and TI values for all video contents. The color points represent the MoTV contents as the black ones represent VT contents. Note that for the VT contents, only the average per scenario (i.e. 10 video sequences) is displayed for clarity reasons. The VT contents belong to categories A and B according to the classification proposed in Annex A of ITU-T Rec. P.910, namely “head-and-shoulders” content with graphics and more details. The values obtained for the VT contents are in accordance with the values shown in P.910, i.e. SI values below 100 and TI values below 50, as the scenes do not exhibit a high spatial and temporal complexity. The contents

Table 2.3 Audiovisual sequences description for MoTV and videotelephony applications

Application	Name	Video	Audio
MoTV	Football	Soccer game	Speech on babble noise from the crowd
MoTV	Movie	Trailer	Speech on music
MoTV	Interview	H&S female	German speech
MoTV	Music	Music clip with singer	Pop music
VT	Birthday	H&S male	German speech
VT	Car reservation	H&S male	German speech
VT	Doctor appointment	H&S female	German speech
VT	Kitchen purchase	H&S female	German speech

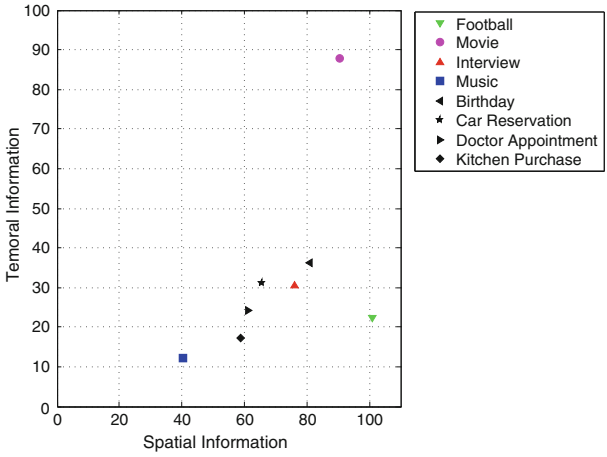

Fig. 2.6 Spatial and temporal information for Mobile TV and videotelephony sequences

for MoTV span a wider range of TI and SI values, notably for the "Football" scene that has a high SI value due to the highly textured football field with players and the "Movie" scene having a high values of both SI and TI as it contains more complex motion like an explosion for instance. Finally, the "Interview" content is close to at least two VT contents showing a clear similarity between the scenes.

2.4.2 Simulated Conversations Scenarios

As described in Sect. 2.3, the audiovisual adaptation of the simulated conversational structures consist of long audiovisual samples that simulate one side of a normal video call conversation. Each of these simulations concerned a unique topic as explained

in Sect. 2.4.1. Each topic is composed of ten sentences, thus leading to 10 samples of approximately 9 s and assigned to a single speaker. The samples are interspersed with a 9 s pause during which test participants are required to answer a question related to the content, in order to "simulate" a real conversation and by doing so distracting the attention of the subject from the quality assessment task. The conversations were divided in two segments of 90 s, thus leading to eight long samples. The content of the scripts used to realize the simulated conversations is reported in Appendix A.

2.5 Interactive Conversational Scenarios

Three different conversational scenarios were used in the interactive experiments (AV-VTi-1, AV-VTi-2 and AV-VTi-3). The first one is referred a "SCT" which stands for short conversation test. The SCT scenarios were developed for use in audio-only conversations [119] and therefore mainly involve the utilization of the audio channel. They were designed to represent real-life telephone conversations, like ordering a pizza, leading to semi-structured dialogues of about 2–3 mins.

The second scenario is the building block scenario ("BB") described in ITU-T Rec. P.920. In this scenario, one conversing partner receives an already assembled item made from colored blocks, as his interlocutor is simply given the spare pieces. The subject with the assembled item has to provide directives to the other subject on how to mount the pieces together. In practice, this scenario consists of showing the item to the camera and providing assistance in case of difficulties. For this scenario, the use of the video channel is essential as it is much easier to simply look at the item on the screen rather than following a string of complicated instructions. These two scenarios can be regarded as making an unbalanced usage of the audio channel for the SCT scenario and of the video channel for the BB scenario. There was thus a need to develop a more balanced type of scenario reflecting everyday usage of videotelephony, much like the SCTs for telephony.

As a consequence, an audiovisual version of the SCTs was developed by adapting the audiovisual simulated conversations to an interactive context. This type of scenario was intending to simulate an "average" videotelephony conversation with a balanced use of the audio and video channels, notably through the use of visual cues necessary for carrying out the conversational task. 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 Rental" scenario translated in english is provided in Table 2.4. This example describes the structure of a conversation between two interlocutors by detailing the temporal sequence in terms of semantic contents (i.e. questions and answers) and physical actions to be completed (e.g. showing a picture to the camera).

This scenario was named audiovisual short conversation test ("AVSCT") as a reference to the SCT scenarios. The script of these scenarios can be found in [62] and is also reported in Appendix B. Two interactive experiments containing the SCT and AVSCT scenarios were carried out, namely AV-VTi-2 and AV-VTi-3, and the BB

Table 2.4 Extract of the conversational scenario "Car Rental"

Interlocutor ID	Action	Content/Instruction
1	Ask	"What kind of vehicle would you like to rent?"
2	Answer	"A break"
2	Do	Show a picture of a break
2	Ask	"How does the small pick-up look like in the offer?"
1	Answer	Describe the small pick-up (color etc.)
1	Do	Show a picture
1	Ask	"When do you want to rent the vehicle?"

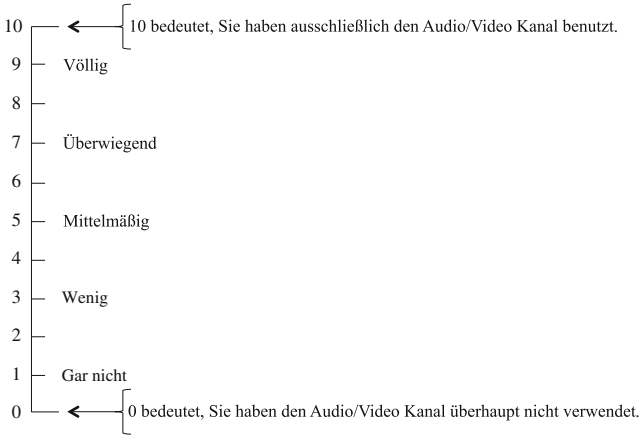

Fig. 2.7 Rating scale for measuring the degree of utilization of the audio and video channels for interactive tasks. The equivalent in the english language of the scale's labels would be: "entirely" for "völlig", "mostly" for "überwiegend", "moderately" for "mittelmäßig", "little/somewhat" for "wenig" and "not at all" for "gar nicht"

scenario was additionally used in the AV-VTi-3 experiment. In both experiments the subjects were asked, after each dialog, to assess to which extent they actually did pay attention to the audio and video channels and evaluate their usage of each channel in order to fulfill the task. They rated each utilization degree using the 11-point scale shown in Fig. 2.7.

At the extremities of the scale (at points 0 and 10), an additional description was added: 0 meaning that the subject hasn't used the channel at all and 10 that the subject has exclusively used this channel. The comparison between the utilization of both audio and video signals rated on an intensity scale for both experiments is depicted in Fig. 2.8a and b for experiment AV-VTi-2 (interactive scenarios SCT and AVSCT respectively) and in Fig. 2.9a, b and c for experiment AV-VTi-3 (interactive scenarios SCT, AVSCT and BB respectively). The graphs report the scores mapped onto a 5-point scale for different test conditions.

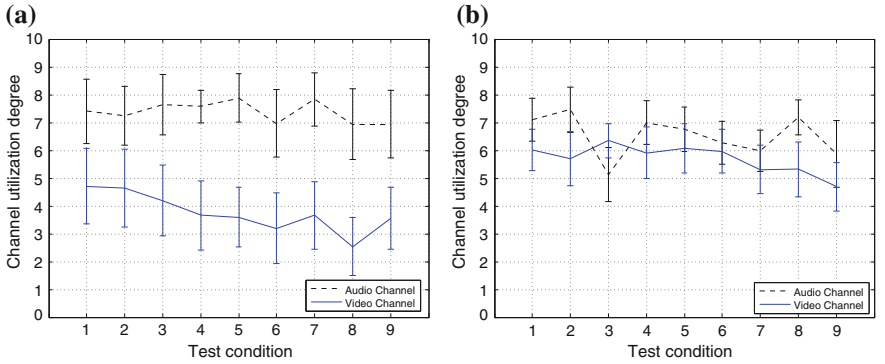

Fig. 2.8 Degree of audio and video channel utilization depending on the conversational scenario for experiment AV-VTi-2. **a** SCT scenario. **b** AV SCT scenario

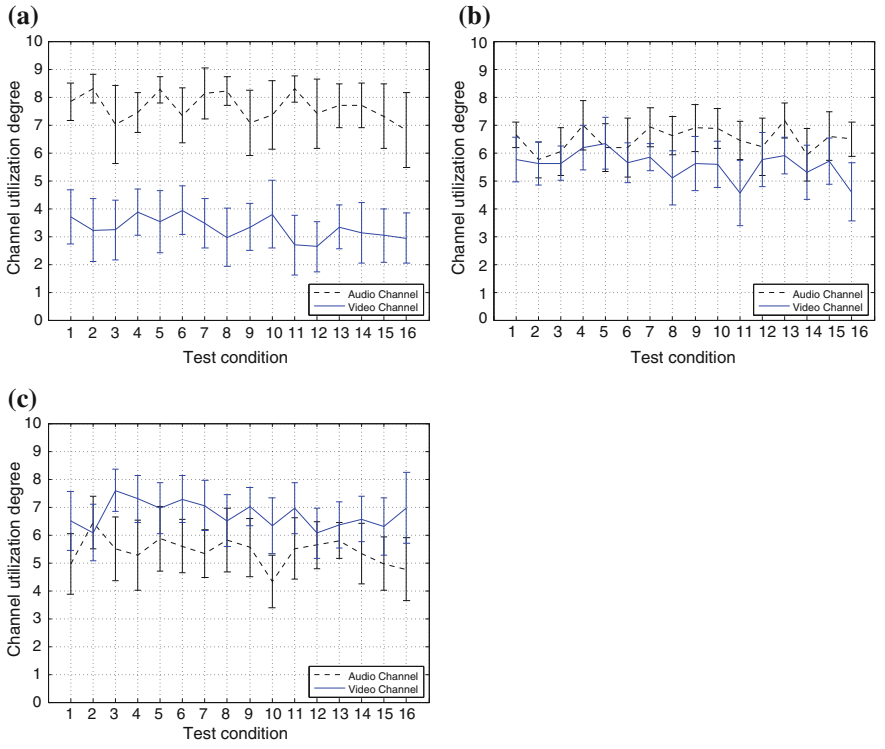

Fig. 2.9 Degree of audio and video channel utilization depending on the conversational scenario for experiment AV-VTi-3. **a** SCT scenario. **b** AV SCT scenario. **c** BB scenario

Table 2.5 Influence of the conversational scenario on the audio and video channel utilization degree. The variable mean audio (resp. video) represents the average degree of utilization of the audio (resp. video) channel over all test conditions

Experiment	Scenario	F	p	Mean audio	Mean video
AV-VTi-2	SCT	196.7	<0.001	7.4	3.7
AV-VTi-2	AVSCT	16.3	<0.01	6.5	5.7
AV-VTi-3	SCT	708.7	<0.001	7.7	3.3
AV-VTi-3	AVSCT	43.3	<0.01	6.5	5.6
AV-VTi-3	BB	61	<0.01	5.4	6.7

The utilization degree of the audio and video channels for the SCT scenario was approximately constant in both experiments, with a much larger use of the audio channel. The means of the audio and video channel usage are significantly different for all conditions in both experiments, see Table 2.5. For the AVSCT scenario, the utilization is more balanced between both channels, although a slight advantage is given to the audio channel as these scenarios are still dialog-based (mean audio: 6.5, mean video: 5.7). For condition 3 which is defined as "excellent" video transmission and "bad" audio transmission, the advantage is given to the video channel (even though the difference is not significant). This reversal can be interpreted as a conscious shift of attention. The utilization degree can possibly change depending on the transmission characteristics. Finally, it can be seen in Fig. 2.9c that the video channel is on an average more used for the Building Block scenario even if the differences are not significant for all conditions (mean audio: 5.4, mean video: 6.7). Consequently, the assumptions made on the scenarios were experimentally validated and revealed that the type of scenario is a factor of influence by driving the attention of the user, which in turn impacts the quality perception.

2.6 Summary

This chapter introduced an experimental approach for the subjective evaluation of audiovisual quality for videotelephony applications, both in a passive and in an interactive context. Several aspects of audiovisual quality will be explored: first the separate evaluation of the audio and video qualities for videotelephony based on short samples; then, characteristics of the audiovisual integration with regard to the type of experimental context (passive or interactive), and the type of interactive scenario; finally, characteristics of temporal integration by linking up the momentary quality of short samples to longer sequences called simulated conversations.

To that end, a test bed was developed that was composed of a videotelephony client and of a control module. Through the control module, the video client can be used in an off-line mode to automatically process audiovisual sequences according to pre-defined test conditions including coding and transmission degradations. A user

graphical interface was created to be used in conjunction with the control unit, to enable real-time video calls experiments.

Moreover, three usage scenarios were foreseen: first, the simulated conversations, initially designed for speech have been adapted to the audiovisual context in order to study the user experience of temporal degradations. These simulated conversations were constituted of several independent short samples that could be used separately as a second usage scenario. Finally, the simulated conversations were adapted to fully interactive situations by modifying their structure into a question/answer type of semi-structured dialog. The semantic content of these scenarios was thus expected to be close and to allow inter-experiment comparisons.

http://www.springer.com/978-3-319-14165-7

Audiovisual Quality Assessment and Prediction for Videotelephony

Belmudez, B.

2015, XVIII, 184 p. 62 illus., 36 illus. in color., Hardcover
ISBN: 978-3-319-14165-7