Perceived quality of audio-visual stimuli containing streaming audio degradations

Helard Martinez
University of Brasilia
Brasilia, Brazil
helardb@unb.br

Mylène C.Q. Farias
University of Brasilia
Brasilia, Brazil
mylene@ieee.org

Andrew Hines
University College Dublin
Dublin, Ireland
andrew.hines@ucd.ie

***Abstract*—Multimedia services play an important role in modern human communication. Understanding the impact of multisensory input (audio and video) on perceived quality is important for optimizing the delivery of these services. This work explores the impact of audio degradations on audio-visual quality. With this goal, we present a new dataset that contains audio-visual sequences with distortions only in the audio component (Im-AV-Exp2). The degradations in this new dataset correspond to commonly encountered streaming degradations, matching those found in the audio-only TCD-VoIP dataset. Using the Immersive Methodology, we perform a subjective experiment with the Im-AV-Exp2 dataset. We analyze the experimental data and compared the quality scores of the Im-AV-Exp2 and TCD-VoIP datasets. Results show that the video component act as a masking factor for certain classes of audio degradations (e.g. echo), showing that there is an interaction of video and audio quality that may depend on content.**

***Index Terms*—QoE, audio-visual quality, VoIP, immersive experimental methodology**

I. Introduction

The prominent evolution of digital communication systems has led to a significant growth of multimedia services in the past two decades. Services like Video on Demand (VoD), Internet protocol Television (IPTV), and video-telephony have an important role in modern human communication and entertainment systems. Tools, that help guarantee a certain level of user's Quality of Experience (QoE), contribute to the success of such services. Objective quality metrics have been developed over the years to predict the perceived quality of different types of media such as audio [1], video [2], and audio-visual content [3]. Their performance is generally measured by computing the correlation of their estimates and of the scores given by participants in experiments.

Psychophysical experiments, in which a number of human participants rate the perceived quality of a content, are considered the most accurate way of measuring quality. These experiments provide valuable data that contributes to the development of more precise objective image quality assessment methods. Recommendations for conducting these experiments have been published by international agencies (e.g. ITU and EBU) with the objective of standardizing the experimental methodologies [4]. Nevertheless, experimental methodologies still encounter difficulties when trying to faithfully estimate quality as perceived by the users, especially when considering more natural scenarios. This has motivated the proposal of new experimental methodologies.

Pinson *et al.* [5] proposed the Immersive Methodology, which puts human participants in a more natural context with the goal of collecting more accurate quality scores. In recent years, several subjective experiments have been carried out using the Immersive Methodology with satisfactory results. For instance, Garcia [6] used the Immersive Methodology to study the quality of long-duration sequences under an adaptive streaming system. Furthermore, Robitza [7] carried out an immersive experiment to study the impact of quality variations and stalling events on audio-visual sequences with durations of over a minute.

In this paper, we used the Immersive Methodology to perform a subjective experiment with the goal of estimating the quality of audio-visual sequences. Quality scores were gathered for a set of audio-visual sequences with distortions only in the audio component. The TCD-VoIP dataset [8] served as a reference to produce a new audio-visual dataset with only-audio distortions: the Im-AV-Exp2. The experiment had the goal of recreating some of the streaming audio degradations from the TCD-VoIP dataset on an audio-visual scenario and analyzing the effect of such degradations on the perceived audio-visual quality. More importantly, the experiments had the goal of testing the effect of the visual content on the overall quality. This experiment is part of a set of subjective experiments investigating the impact of audio and video degradations on the overall audio-visual quality. Findings from these experiments will be used to analyse the relationship between streaming and compression artifacts on audio-visual quality.

The rest of this paper is organized as follows. In Section II, a brief description of the TCD-VoIP dataset is presented. Section III describes the subjective experiment performed in this work, including the experimental methodology, source stimuli (Im-AV-Exp2 dataset), and test conditions. Section IV presents the experimental results, while Section V presents a comparison of the quality scores of the TCD-VoIP and Im-AV-Exp2 datasets is presented. Finally, Section VI presents the conclusions.

II. The TCD-VoIP Dataset

The TCD-VoIP dataset contemplates a number of degradations that commonly occur in a voice over IP transmission environment. Such degradations are referred to as platform-independent, that is, degradations that occur independently of the codec, hardware, or network in use [8]. The dataset

ISBN 978-90-827970-1-5

contains five types of degradations: 1) background noise, 2) competing speakers, 3) echo effects, 4) amplitude clipping, and 5) choppy speech. For each type of degradation, a number of test conditions is set. These test conditions are applied to a set of speech samples, resulting in the TCD-VoIP dataset.

Detailed information regarding the TCD-VoIP database and the subjective results can be found in the work by Harte *et al.* [8]. A brief description of the types of degradations and their parameters is presented next. It is worth mentioning that for the present experiment only four degradation types were used (background noise, choppy speech, clipping, and echo).

A. Background Noise

Background noise is described as any sound other than the sound being monitored. The TCD-VoIP database uses four types of common noise: car noise, road noise, speech babble noise, and office noise.

B. Chop Speech

This type of degradation consists of speech signals in which samples are missing. In a VoIP context, the loss of speech samples might be produced by hardware overload during a VoIP call. To simulate a choppy speech effect, random samples were discarded from the original signal. Missing samples were treated according to three criteria (or chop modes with missing samples): 1) replaced with zeros, 2) completely deleted, and 3) overwritten with previous samples. Additionally, samples were discarded with different lengths (Chop Period) and frequencies (Chop Rate).

C. Clipping

Clipping consists of attenuating the incoming signal amplitude to maintain it below the maximum level permitted. As a result, samples become clipped, introducing distortions in the transmitted signal. Clipping effects were simulated by multiplying the samples amplitude by a constant term.

D. Echo

The TCD-VoIP database uses an echo scenario where copies of the signal being transmitted are picked up at the receiving microphone and then added to a returning signal. To simulate an echo effect, delayed versions of the signal at different SNR values were added to the original signal. Three parameters were varied: 1) Echo alpha: the percentage amplitude of the first delayed version with respect to the original, 2) Echo delay: the delay between the first delayed version of the signal and the original, and 3) Feedback factor: the percentage reduction applied to each of the following delayed versions of the signal.

III. Subjective Experiment Description

A. Methodology

Pinson's Immersive Methodology [5], which has been successfully applied on several subjective experiments [6], [7], was used for designing this experiment. The immersive approach presented by Pinson aims to place the human participant in a more natural scenario in order to capture a more realistic estimate of the perceived quality. A natural scenario is referred here as a recreation of the same conditions in which users generally consume the multimedia service. We used the Immersive Methodology to set three aspects of our subjective experiment: 1) length of stimuli, 2) content diversity and 3) input media.

To engage the participant and put him/her in a more realistic *media-consumption* scenario, the Immersive Methodology recommends the use of longer stimuli sequences (e.g. 30-60 seconds). This recommendation is based on the fact that short-duration sequences (e.g. 8-10 seconds) are perceived as artificial and are unable to transmit an entire idea.

Traditional methodologies often use a reduced pool of source stimuli and process them with a high number of HRCs, deriving a large stimuli pool of repeated content. Observing/listening a large number of sequences with a repeated content leads to fatigue and content memorization, which indefectibly affects the quality of the responses of the participants. The immersive design aims to prevent fatigue and content memorization by presenting each source stimuli only once. That is to say, all participants will rate every HRC, but they will not rate every sequence from the test stimuli pool.

Regarding the input media, the Immersive Methodology recommends to use audio-visual sequences (audio + video components) as stimuli source. This recommendation is made on the grounds that using a video-only (or audio-only) stimuli does not reflect the ways users consume a multimedia service, since a transmission of a video without sound is not likely to occur [9]. Using audio-visual sequences as test material requires that participants rate the overall audio-visual quality, independent of the component under study.

B. Source Stimuli

For this experiment, forty (40) high-definition sequences (video with accompanying audio) served as source stimuli material (SRC). All sequences were pre-processed to standardize main audio and video characteristics. For the audio component, the bit-depth and sample frequency were set to 16 bits and 48 kHz, respectively. As for the video component, the spatial and temporal resolutions were fixed at 1280x720 (720p) and 30 frames per second (fps), respectively, while the color space format was set to 4:2:0.

With regard to the content, the researchers selected a group of sequences considered to be "interesting" and "relevant", especially with respect to the audio component of the sequence (given the experiment's nature). Our goal was to cover different categories of media entertainment. Based on their audio content, source sequences were organized and then labeled as: Sports, Music, Movies, and Speech.

The length of the source sequences varied between 19 and 54 seconds, with an average duration of 35 seconds. Special care was taken to guarantee that the resulting SRC sequences presented a complete idea. That is, in order to maintain the naturalness of the sequences, no cuts were made in the middle of a speech or statement and no scenes were abruptly interrupted.

ISBN 978-90-827970-1-5 © EURASIP 2018

TABLE I
HRC AND ANCHOR TEST CONDITIONS (ANC) PARAMETERS USED IN IM-AV-EXP2.

BG Noise	Noise	SNR (dB)	
HRC1	car	15	
HRC2	babble	10	
HRC3	office	10	
HRC4	road	5	
ANC1	-	-	
Chop	**Period (s)**	**Rate (chops/s)**	**Mode**
HRC5	0.02	1	previous
HRC6	0.02	2	zeros
HRC7	0.04	2	previous
HRC8	0.02	5	zeros
ANC2	-	-	-
Clipping		**Multiplier**	
HRC9		11	
HRC10		15	
HRC11		25	
HRC12		55	
ANC3		-	
Echo	**Alpha (%)**	**Delay (ms)**	**Feedback (%)**
HRC13	0.5	25	0
HRC14	0.3	100	0
HRC15	0.175	140	0.8
HRC16	0.3	180	0.8
ANC4	-	-	-

C. Test Conditions

As mentioned before, four common streaming types of degradations were considered for this particular experiment. The Im-AV-Exp2 dataset was built following the same processing method used in the TCD-VoIP dataset. For each type of degradation, four test conditions were selected from the TCD-VoIP dataset and presented as a particular Hypothetical Reference Circuit (HRC). These test conditions were selected empirically with the goal of covering the entire range of quality observed in the TCD-VoIP dataset. As a result, sixteen (16) HRC arrangements were obtained. The HRCs were organized according to the type of degradation. Additionally, one test condition without degradations was used as an anchor (ANC) to help participants establish the range of quality used in the experiment. Details of the HRCs (HRC1 to HRC16), their corresponding parameters, and the anchor test conditions (ANC1 to ANC4) are presented on Table I.

D. Apparatus and Physical Conditions

This experiment was conducted in a recording studio with sound isolation with one participant at a time. The hardware set-up consisted of a desktop computer, a LCD monitor, and a set of earphones. Additionally, a dedicated sound card (Asus Xonar DGX 5.1) was used to provide participants with an optimal sound experience (in terms of hardware). Participants were seated straight ahead of the monitor with a distance of three screen heights between their eyes and the monitor screen, as suggested by the ITU-T Recommendation BT.500.1.

The experiment was performed with 40 participants (15 female and 25 male) aged from 21 to 36. They were considered naive on digital video and audio defects and their associated terminology. No vision or hearing tests were performed, unimpaired hearing was a pre-requirement, moreover, participants were asked to use glasses or contact lenses if they needed them to watch TV. Additionally, the volume level was set at the middle of the volume bar.

E. Experimental Procedure

The entire experiment was divided into three sessions: Display, Training, and Main sessions. The Display session presented to the participants a set of sample sequences containing all test conditions (HRCs) for each type of distortion (noise, chop, clip, and echo). This session had the goal of showing to the participants the entire quality range of the test sequences. At the end of this session, participants were asked if they noticed quality difference between the sequences.

During the Training session, participants were presented with two sample sequences. After each sequence was played, participants were asked to rate their quality. The objective of this session was to familiarize participants with the data-entry procedure used in the Main session. To rate the sequences quality, participants were presented with a five point Absolute Category Rating (ACR) scale, ranging from '1' to '5'. The scale was labeled as "Bad", "Poor", "Fair", "Good", and "Excellent". In the Main session, participants completed the actual experimental task.

F. Quality Score Measurement

Responses given by participants in a psychophysical experiment are called subjective scores. Traditionally, the mean opinion score (MOS) for each test video is obtained by taking the average of the subjective scores given by all observers to each test sequence. In our experiment, the Mean Quality Score (MQS) per-HRC is obtained by averaging the quality scores, given by all subjects, to the j-th HRC:

$$\mathrm{MQS}_{\mathrm{HRC(j)}} = \frac{1}{n} \cdot \sum_{i=0}^{n} QS_j(i), \tag{1}$$

where n is the total number of subjects and $QS_j(i)$ is the quality score given by the i-th subject to the j-th HRC test sequence, with $j = \{1, 2, \ldots, 16\}$. In other words, $\mathrm{MQS}_{\mathrm{HRC(j)}}$ gives the average quality score for the j-th HRC, measured over all subjects and originals.

IV. EXPERIMENTAL RESULTS

This section presents the analysis of the degradation conditions (i.e. Echo, Chop, Clip, Noise), which are considered service aspects that may be affected during streaming. Figure 1 presents the *Mean Quality Score* (MQS), including a 95% confidence interval, for all HRCs corresponding to the four audio distortions. Results are grouped according to the corresponding audio distortion type.

For the Background Noise distortion type, each HRC corresponds to a combination of a noise type and an SNR value, as detailed in Table I. It can observed that the quality scores rarely reach 3 points in the MQS scale. These results are in accordance with previous results that showed that, for noise SNR values bellow 20dB, the quality scores are around 3 points or less [8]. Analyzing the parameters, it can be observed that sequences with an SNR value below 15dB obtained quality scores smaller than 3 points. For the particular case of HRC2 and HRC3, which both present the same SNR value (10dB), it can be observed that the babble noise was more annoying than the office noise. Such behavior is again in accordance with results from previous (audio-only) experiments [8].

ISBN 978-90-827970-1-5 © EURASIP 2018

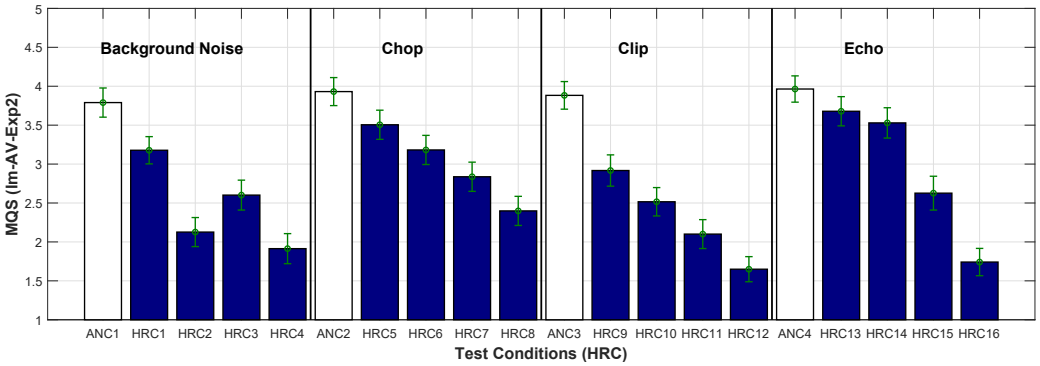


Fig. 1. MQS for all four distortions. See HRC specifications in Table I.

For the Chop distortion, each HRC corresponds to a combination of three parameters (rate, period, and mode), as detailed in Table I. It can be noticed that the MQS values vary from 2.5 to 3.5, with the MQS values decreasing from HRC5 to HRC8. This behavior seems to be closely related to the chop rate value. An analysis of the parameters suggests that the perceived quality decreases as the chop rate increases, independently of the chop mode. In particular, for a fixed rate of 2 chops/second, repeating previous portions of samples (*previous* mode) is slightly more annoying than inserting silence portions (*zeros* mode).

For the Clip distortion type, the MQS values vary between the 3 and 1.5 points decreasing from HRC9 to HRC12. Such results might suggest that clipped distortions are perceived as more severe. For the particular case of HRC9 and HRC10, where the multipliers values are 11 and 15, respectively, quality scores below the 3 points are observed. These results are particularly interesting since previous (audio-only) experiments found similar quality scores for multiplying factors above 18 [8].

For the Echo distortion type, each HRC corresponds to a combination of three parameters (alpha, delay, and feedback), as detailed in Table I. The MQS values vary between 3.7 and 1.7 points in the MQS scale. Although the HRCs quality values decrease from HRC13 to HRC16, an abrupt drop in MQS is observed between HRC14 and HRC15. For this particular case, it can be observed that the presence of a feedback affects considerably the perceived quality. These results were also observed in previous audio-only studies, where the inclusion of a feedback produced the lowest quality scores [8].

V. Comparison of Datasets

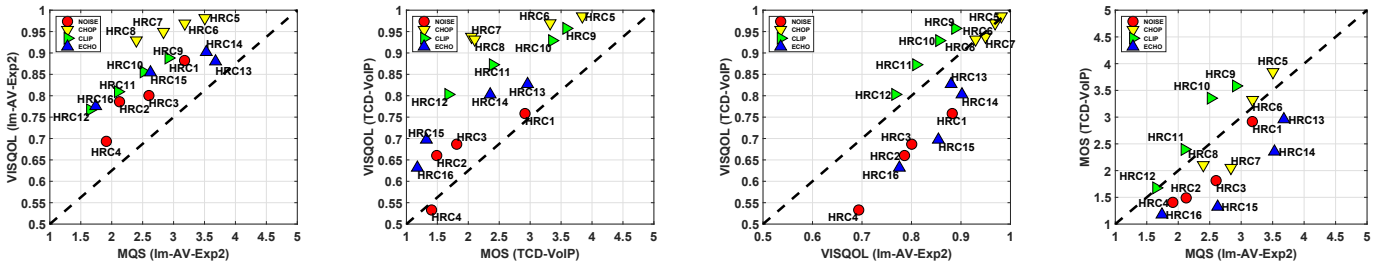
In this section, we compare the objective and subjective quality responses for both datasets. It is worth pointing out that there are obvious differences between the two datasets. First, the TCD-VoIP dataset contains only speech audio sequences, while the Im-AV-Exp2 dataset contains speech, sport, movies, and music audios in audio-visual sequences. Second, the two datasets used different experimental methodologies to collect the subjective scores. Despite these differences, a comparison of theses two datasets can provide interesting insights regarding the impact of the visual component on the overall quality perception, when the stimuli contains streaming degradations (only) in the audio component.

To perform this comparison, we used two versions of an objective quality metric to establish a similar measure for both datasets. In TCD-VoIP, the VISQOL speech model [10] was used to estimate the speech quality of the stimuli. Meanwhile, in Im-AV-Exp2, the VISQOLAudio quality metric [1] was used to obtain the quality of the audio component of the stimuli. Then, we compared the subjective quality scores, **MQS (Im-AV-Exp2)** and **MOS (TCD-VoIP)**, of both datasets with the corresponding VISQOL objective scores, **VISQOL (Im-AV-Exp2)** and **VISQOL (TCD-VoIP)**. Figure 2 depicts scatter-plots showing comparisons of these objective and subjective scores.

Figures 2 (a) and (b) show the subjective scores versus the VISQOL scores for the Im-AV-Exp2 and TCD-VoIP datasets, respectively. Notice that the VISQOL metric tends to overestimate the quality for all degradations in both datasets. Interestingly, we observe that VISQOL ranked all degradations in the same order for both datasets, i.e. Chop degradations were rated as less annoying, followed by Clip, Echo, and Noise degradations. These results show that the characteristics of the audio degradations seem to be affecting the perceptual quality of the stimuli of both datasets in a similar way.

Figure 2 (c) depicts a scatter-plot of the VISQOL scores for TCD-VoIP versus the VISQOL scores for Im-AV-Exp2. From the plot in this figure, we can notice that, in the both datasets, there is a consistency of the results corresponding to the Chop degradations (identical results would correspond to points in the diagonal traced line). It is worth pointing out that the VISQOL scores for the Chop degradations had high values (over 0.9 for both datasets). Regarding the Clip degradations, the VISQOL scores obtained for the TCD-VoIP dataset were higher than the VIQOL scores obtained for the Im-AV-Exp2 dataset (i.e. points are above the diagonal line). The VISQOL scores for Echo and Noise degradations, on the other hand, were smaller for the TCD-VoIP dataset than for the Im-AV-Exp2 dataset (i.e. points below the diagonal line). This result shows that, although the sample conditions for both datasets were generated using the same technique, the content had an influence on the perceived quality, causing a quality increase

ISBN 978-90-827970-1-5 © EURASIP 2018



(a) (b) (c) (d)

Fig. 2. Subjective-Objective comparison for Im-AV-Exp2 and TCD-VoIP.

(Clip) or decrease (Chop and Noise) for speech content (TCD-VoIP) to general audio content (Im-AV-Exp2).

Figure 2 (d) depicts a scatter-plot of subjective scores for TCD-VoIP versus the subjective scores for Im-AV-Exp2. The comparison is made between the audio-only quality scores, MOS(TCD-VoIP), and the corresponding audio-visual scores, MQS(Im-AV-Exp2). Notice that, although these quality scores come from different experiments with different content and different conditions, there is again a consistency between the audio-only and audio-visual scores for the Clip and Chop degradations, with only a few exceptions are far from the diagonal line (HRC7, HRC9, and HRC10). The subjective scores for the Noise degradations lie below the diagonal line, but not too far from it. It is interesting to note that, for the Echo degradations, the audio-only subjective scores are consistently higher than the audio-visual scores (i.e. points are below the diagonal line). This suggests that the video component has a more pronounced impact for Echo degradations, acting as a masking factor and producing higher quality scores. In other words, the Echo degradation had a smaller impact on the perceived overall quality of audio-visual stimuli than on the perceived audio quality of of audio-only stimuli. This result seems to be in agreement with previous studies [11] where participants rated echo distortions as imperceptible during video calls, i.e., in the presence of a visual component.

VI. Conclusions

This paper presents the results of a subjective experiment conducted with an audio-visual dataset containing distortions only in the audio component of the stimuli. The degradations of the dataset consisted of commonly encountered audio streaming degradations, which matched those found in the audio-only TCD-VoIP dataset. Results were in accordance with previous experiments that used the same type of distortions. Moreover, it seems that participants were able to distinguish between the different levels of quality for each type of degradation. It is worth poiting out that the Im-AV-Exp2 dataset is part of a wider audio-visual database, which contains combinations of audio and video degradations.

We also compared the quality scores of Im-AV-Exp2 and TCD-VoIP. The VISQOL Audio quality metric was used to establish a ground common measure for the comparison. Different degradations showed the same general trend of behavior in both datasets. Moreover, results showed that the video component acted as a masking factor for certain classes of audio degradations, e.g. echo. These results help to understand how audio streaming degradations are affected by the visual component, and this knowledge can be used in the design of real-time multimedia QoE metrics.

Acknowledgment

This work was supported by the Conselho Nacional de Desenvolvimento Cientfico e Tecnológico (CNPq), the Coordenação de Aperfeiçoamento de Pessoal de Nível Superior (CAPES), and the University of Brasília (UnB).

References

[1] A. Hines, E. Gillen, D. Kelly, J. Skoglund, A. Kokaram, and N. Harte, "ViSQOLAudio: An objective audio quality metric for low bitrate codecs," *The Journal of the Acoustical Society of America*, vol. 137, no. 6, pp. EL449–EL455, 2015.

[2] M. Vranješ, S. Rimac-Drlje, and K. Grgić, "Review of objective video quality metrics and performance comparison using different databases," *Signal Processing: Image Communication*, vol. 28, no. 1, pp. 1–19, 2013.

[3] Z. Akhtar and T. H. Falk, "Audio-Visual Multimedia Quality Assessment: A Comprehensive Survey," *IEEE Access*, vol. 5, pp. 21 090–21 117, 2017.

[4] ITU-R, "Recommendation BT.500-8: Methodology for subjective assessment of the quality of television pictures," International Telecommunication Union, Geneva, Tech. Rep., 1998.

[5] M. Pinson, M. Sullivan, and A. Catellier, "A New Method for Immersive Audiovisual Subjective Testing," in *Proceedings of the 8th International Workshop on Video Processing and Quality Metrics for Consumer Electronics (VPQM)*, 2014.

[6] M.-N. Garcia, D. Dytko, and A. Raake, "Quality impact due to initial loading, stalling, and video bitrate in progressive download video services," in *Quality of Multimedia Experience (QoMEX), 2014 Sixth International Workshop on*. IEEE, 2014, pp. 129–134.

[7] W. Robitza, M. N. Garcia, and A. Raake, "At home in the lab: Assessing audiovisual quality of HTTP-based adaptive streaming with an immersive test paradigm," in *Quality of Multimedia Experience (QoMEX), 2015 Seventh International Workshop on*. IEEE, 2015, pp. 1–6.

[8] N. Harte, E. Gillen, and A. Hines, "TCD-VoIP, a research database of degraded speech for assessing quality in VoIP applications," in *Quality of Multimedia Experience (QoMEX), 2015 Seventh International Workshop on*. IEEE, 2015, pp. 1–6.

[9] J. G. Beerends and F. E. De Caluwe, "The influence of video quality on perceived audio quality and vice versa," *Journal of the Audio Engineering Society*, vol. 47, no. 5, pp. 355–362, 1999.

[10] A. Hines, J. Skoglund, A. C. Kokaram, and N. Harte, "ViSQOL: an objective speech quality model," *EURASIP Journal on Audio, Speech, and Music Processing*, vol. 2015, no. 1, p. 13, 2015.

[11] M.-D. Cano and F. Cerdan, "Subjective QoE analysis of VoIP applications in a wireless campus environment," *Telecommunication Systems*, vol. 49, no. 1, pp. 5–15, 2012.