Audiovisual Quality Estimation for Video Calls in Wireless Applications

Mohammad Goudarzi, Lingfen Sun, Emmanuel Ifeachor
School of Computing and Mathematics, University of Plymouth, UK
Email: {mgoudarzi, lsun, eifeachor}@plymouth.ac.uk

Abstract—Mobile/wireless multimedia applications (e.g., video calls and IPTV) have gained great momentum in recent years. An important issue is to monitor/predict overall audiovisual quality, instead of audio-only or video-only quality, non-intrusively for technical or commercial reasons. Previous audiovisual modeling research mainly considered application parameters (e.g., codec and send bit rate). Little attention has been paid to how network parameters, e.g., Packet Error Rate (PER) affect audiovisual quality. The aim of this paper is to explore methods to predict audiovisual quality objectively for video calls in wireless applications. The contributions of the paper are twofold. Firstly, we present subjective test results on how audio and video contribute to overall audiovisual quality and develop models to reflect this relationship. Secondly, we investigated how network parameters (e.g., PER) and application parameters, e.g., video Frame Rate (FR) affect overall audiovisual quality. We developed a regression model to predict audiovisual quality from PER and FR which can be used to monitor/predict audiovisual quality non-intrusively. We also explore the possibility to predict audiovisual quality from full-reference voice and video quality metrics (i.e., PESQ and PSNR) and from combined PESQ/PSNR and network/application parameters. The different prediction accuracy obtained from these models (accuracy from 84% to 93%) indicates the complex attributes in audiovisual quality prediction. An extended Evalvid/NS-2 platform is developed to support simulation of video calls over wireless networks.

Keywords- Audiovisual quality, quality assessment model, non-intrusive, subjective tests, H.263, G.711, NS-2

I. INTRODUCTION

Multimedia systems are becoming more and more important and video calls in mobile and wireless networks are becoming a standard part of user’s experience. Thus, it is essential for network operators to provide good levels of audiovisual quality in order to maintain customer satisfaction. Research on Quality of Experience (QoE) for multimedia services in communication systems has traditionally focused on individual modalities (e.g. video or audio). Significant progress has been made in developing perceptual objective measurement models for audio [1] and video [2] respectively. Limited effort has been made to develop audiovisual quality models based on both audio and video qualities.

There are two fundamental aspects that need to be considered in order to develop a reliable quality metric for audiovisual content: 1) the perceptual process for the overall audiovisual quality. It is important to be able to measure and further understand the interactions between the Audio Quality (AQ) and Video Quality (VQ) in order to accurately model the overall AudioVisual Quality (AVQ) as perceived by the end-user. 2) the effects of transmission errors encountered in wireless and mobile networks (e.g., packet error rate or link error rate) and application related parameters (e.g., codec and frame rate) on user’s perception of audiovisual quality.

Various studies have shown that in order to measure the quality of the multimedia content effectively, both the video and the audio must be taken into consideration [3]–[7]. Several audiovisual evaluations have shown that the AQ considerably affect the overall assessment of the audiovisual content [4] and VQ ratings given by subjects improved with better audio quality [7]. A number of studies have focused on the understanding of human perceptual processes for audiovisual quality. In [3], [5], [6] the integration of audio and video quality by human subjects was studied and models were proposed for a wide range of audiovisual content. It was shown that for the video-conferencing material, audio quality has a significant impact on the overall audiovisual quality. Studies on audio and video quality interactions such as [5] have focused on low-bitrate video applications scenarios (e.g., mobile video), other studies like [8] developed models based on spatial and temporal properties of the video and Auditory Distance(AD) for audio quality. However, few studies have been carried out to find the effects of transmission errors encountered on low-bitrate, low-motion content in mobile and wireless network scenarios. Other studies have focused on the impact of factors such as frame rate [9] and packet loss [10] on the perceptual quality of multimedia content. But they are only focused on video quality, IPTV and video streaming applications, without consideration of overall audiovisual quality. In this paper, we focus on video call scenarios and the impact of network and applications parameters on the overall audiovisual quality.

In objective audio and video quality measurement, PESQ and PSNR have been widely used to assess voice and video quality. But it is unclear how PSNR and PESQ values relate to overall audiovisual quality, or whether it is possible to derive overall audiovisual quality from PESQ and PSNR values.

In this paper we study the relationships between the two modalities in wireless and mobile environments and the effects of network/application parameters on the overall audiovisual quality in video calls. Our goal is to develop models based on the effects of network parameters, such as Packet Error Rate (PER), and application parameters, such as frame rate (FR) on overall audiovisual quality. We carried out subjective tests to assess audiovisual, audio-only, and video-only quality. Using subjective data, we analyzed the influence of video and audio...
on overall audiovisual quality and proposed models for audiovisual quality using interactions between audio and video. We also investigated the impact of PER and FR on audiovisual quality and proposed a model for predicting perceived audiovisual quality from network/application parameters. Finally, we proposed models for estimating audiovisual quality based on objective audio quality (PESQ), video quality (PSNR) and combined PESQ/PSNR, PER and FR parameters.

The rest of the paper is organized as follows. Section II describes the testbed setup and test material selection. Section III describes the subjective test procedure and the assessment method. Section IV analyzes the experimental results and describes the modeling of audiovisual quality based on the audio/video interaction, subjective and objective data. Section V concludes the paper.

II. TESTBED SETUP

A. Simulation Platform

The Evalvid tool-set [11] is a popular framework among researchers for evaluation of the quality of video transmitted over a real or simulated communication network. Although the new versions of Evalvid can also be used for audio transmission, it only supports AAC audio and MPEG encapsulation of the audio and video. Evalvid also uses hint tracks for packetizing and transmitting the video, which makes it more suitable for simulating video streaming rather than on-the-fly encoding of audio and video in separate RTP streams, as utilized by most video call applications (e.g., x-lite, IMS-Communicator and other popular IM clients). This issue was addressed in [12], by using non-hinted tracks for video and audio evaluation. However, we evaluated this tool-set and found that it uses RTP transport streams, which is not typically used in video call applications. Moreover, it assumes a constant packet size of 1328 for all the packets in a session, which is not a realistic assumption for video calls. This motivated us to build up a new tool-set based on the Evalvid framework that can evaluate audio and video in video call scenarios.

We combined Evalvid/NS-2 with Java Media Framework (JMF), and RTPtools into a new evaluation tool-set for video and audio transmission in separate streams. We use three code snippets called AVTransmit, AVReceive and Export to transmit, receive and save the audio and video RTP streams. RTPtrace is used to parse RTP packets and generate sender trace files and RTPplay can play back the captured RTP sessions. The system architecture of this tool-set is presented in Fig. 1. To simulate a video call, we first transmit the original audiovisual sequences using AVTransmit and extract the information of the audio and video streams to obtain trace files using Wireshark and RTPTools. We then use the traffic trace files as an input to NS-2 as seen in Fig. 1(b). After simulation, we obtain the receiver trace file, that is used to find the location of lost packets. Using both sender and receiver trace files, we can obtain distorted audiovisual sequence through RTPplay, AVreceive and Export applications.

In order to measure the complexity and movement of the test scenes, we used Spatial perceptual Information (SI) and Temporal perceptual Information (TI) as suggested in [13]. SI (Eq. 1) is based on the Sobel edge detection filter, applied to each luminance frame $F_n$ at time instance $n$. TI (Eq. 2) is based on the motion difference feature $ΔF_n$ of every pixel $F_n$ of the luminance plane:

$$SI = \max_{time} \{\text{std}_{space}[\text{Sobel}(F_n)]\}. \tag{1}$$

$$TI = \max_{time} \{\text{std}_{space}[ΔF_n]\}. \tag{2}$$

![Fig. 1. (a) Block diagram of the simulation setup (b) the NS-2 environment](image-url)
SI and TI ranges for our 6 test sequences are illustrated in Fig. 2. It can be seen that the we have used test sequences with low TI that cover a wide range of SI, which are representative of low-motion and low-medium coding complexity of video call scenarios.

![Fig. 2. Spatial(SI) and Temporal(TI) Information of test scenes](image)

C. Audiovisual Test Samples

In this paper we have used H.263 video encoder and G.711/μlaw voice codec for their low complexity and popularity among video-conferencing applications and SIP clients such as x-lite and IMS-communicator. We used a wireless network scenario similar to [2] in our network simulations for this experiment. Packet losses occur in the wireless segment of the network using a Gilbert-Elliot(GE) model [14]. 60 total clips were used for each subjective test session. In our experiments, PER is set to 0.01, 0.05, 0.1, 0.15 and 0.20, and FR is set to 8 and 15.

### III. SUBJECTIVE ASSESSMENT OF AUDIOVISUAL QUALITY

A. Assessment Method

Subjective tests were performed according to ITU-T Recommendations [15], [13] and [16] for audiovisual, video and audio, respectively. We used Absolute Category Rating (ACR) for our experiments using a discrete 9-level quality scale as suggested in [15] for low-bitrate evaluations. Full text instructions were given to the subjects prior to the test and at the beginning of each session.

B. Subjects

To evaluate the subjective perceptual audiovisual quality, we worked with 48 paid participants (26 male and 22 female). The test was divided into audio-only, video-only and audiovisual parts and 16 subjects participated in each part of the test as summarized in Table II. The participants were all staff and students of the University of Plymouth. Some subjects participated in all three parts of the subjective test. Their age ranged between 18 to 40 years. All subjects reported that they had normal or corrected vision and normal hearing.

C. Test Procedure

The subjective tests were carried out in a multimedia lab under the supervision of the research investigator. We created a subjective testing website [17] in order to present the samples to the subjects and to collect the opinion scores of the subjects. The website contains full instructions on how to do the test and a warm-up page at the beginning of each session. The subjective test consisted of 3 sessions of about 1 hour. Test sessions were conducted on 3 consecutive days as listed in Table II. Samples were presented to the subjects in groups of 15 samples and there was a 3-5 minutes pause after every group of samples. There was a total of 4 groups in each session.

### TABLE II

<table>
<thead>
<tr>
<th>Day</th>
<th>Content</th>
<th>Result</th>
<th>Clips</th>
<th>Subjects</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Audio-only</td>
<td>MOSₐ</td>
<td>60</td>
<td>9 male, 7 female</td>
</tr>
<tr>
<td>2</td>
<td>Video-only</td>
<td>MOSᵥ</td>
<td>60</td>
<td>8 male, 8 female</td>
</tr>
<tr>
<td>3</td>
<td>Audiovisual</td>
<td>MOSₐᵥ</td>
<td>60</td>
<td>9 male, 7 female</td>
</tr>
</tbody>
</table>

D. Subjective Test Results

MOSₐᵥ is shown as a function of PER and FR in Fig. 3 for each sample. An analysis of the subjective data reveals that the source material has a big influence on perceived quality. Generally, when the scene had a higher complexity (SI) the VQ is lower than the AQ and lower scores were given for the overall AVQ. It can also be seen that almost all sequences with 15 FR had higher opinion scores given by the subjects.

![Fig. 3. MOSₐᵥ as a function of PER and FR (FR 8/s: △; FR 15/s: ○)](image)

Fig. 4 shows how pairs of audio and video quality levels interact in influencing the overall audiovisual quality. It can be seen that for the same vide quality levels, improving the audio quality generally results in better audiovisual ratings. Also the impact of audio quality at good video quality level is very significant. Increasing audio quality from fair to good has a sharp effect on average overall audiovisual quality ratings when video quality is at good level, compared to other VQ levels.

![Fig. 4. MOS level interaction plot](image)
IV. DATA ANALYSIS AND QUALITY MODELING

A. Audio-Video Interactions

We used Principal Component Analysis (PCA) to study the influence of AQ and VQ and their multiplicative interaction term AQ-VQ [3] on AVQ.

1) Principal Component Analysis: We constructed four-dimensional test vectors composed of MOS_A, MOS_V, MOS_AV and multiplicative interaction MOS_A · MOS_V. Fig. 5(a) shows the eigenvalues corresponding to the four principal components. The first two components account for more than 90% of the variance and hence were used for the modeling of the data. The PCA results from Fig. 5(b) show the influence of individual modalities and the multiplicative interaction term on the overall audiovisual quality. The PCA results provide evidence that both AQ and VQ contribute to AVQ. It can be observed that AQ-VQ also has a significant effect on AVQ.

![Fig. 5. Principal Component Analysis](image)

2) Modeling: In this section, we further investigate AQ and VQ relationship in terms of modeling and prediction. As the PCA results suggest, AVQ can be modeled using all vectors (i.e., AQ, VQ, and also AQ-VQ). We applied stepwise linear regression models to assess the relationship between AQ/VQ and the overall audiovisual quality. The general model is assumed as follows [5] [18]:

\[
\text{MOS}_\text{AV} = \alpha_0 + \alpha_1 \text{MOS}_A + \alpha_2 \text{MOS}_V + \alpha_3 \text{MOS}_A \cdot \text{MOS}_V.
\]  

(3)

The correlation results of regression analysis are summarized in Table III. The results show that using the additive model (AQ+VQ) provides the best fit:

\[
\text{MOS}_\text{AV} = 0.565 + 0.260 \text{MOS}_A + 0.476 \text{MOS}_V.
\]  

(4)

From our results, AQ+VQ achieves the highest accuracy, but adding AQ-VQ did not bring any further improvements to the model. A relatively good modeling is also possible by using only the multiplicative term, however, the accuracy is slightly lower than the additive model. As our audiovisual contents also include the impact from network packet loss, the relationship between audiovisual and audio-only/video-only yield similar results with Winkler’s work [5] for audiovisual quality under different application parameters. Our results also did not follow Hand’s model [3] which indicates a very high importance for the multiplicative component.

B. Regression-Based Audiovisual Quality Prediction

In this section we describe the regression-based audiovisual quality prediction model based on the application level parameter, frame rate (FR), and network level parameter, packet error rate (PER).

Subjective results show that the quality of audiovisual varies at the different values of PER and FR and both MOS_A and MOS_V contribute to the audiovisual quality, with MOS_A having more weight. PER affects both AQ and VQ and FR only affects VQ. The best fit equation based on PER and FR is shown in Eq. 5.

\[
\text{MOS}_\text{AV} = \frac{\beta_0 + \beta_1 \text{FR}}{1 + \beta_2 \text{PER}^2}.
\]  

(5)

The coefficients were obtained by a non-linear regression of the proposed model with our training set (70% of MOS values from subjective tests). The coefficients for Eq. 5 are given in Table V.

We further investigated the interaction of objective measurements of audio and video quality measured by PESQ and PSNR on the overall audiovisual quality obtained from subjective test. AQ_PESQ values are measured using PESQ-LQO [19] and are in the range of 0 to 5. VQ_PSNR values are measured using PSNR in (db). We first derived models based on various fittings of PESQ and PSNR values. The fitting accuracy based on objective measurements are summarized in Table IV.

![Table III: Model accuracy of different fits](image)

An additive model based on PESQ and PSNR provides the highest accuracy among different fittings:

\[
\text{MOS}_\text{AV} = -0.101 + 0.405 \text{AQ}_{\text{PESQ}} + 0.050 \text{VQ}_{\text{PSNR}}.
\]  

(6)

The lower accuracy of this model, compared to the previous model (Eq. 4 with accuracy of 88.09%), can be explained in terms of the correlation between PESQ/PSNR and MOS_A/MOS_V. The correlation results (0.73 for PESQ and 0.643 for PSNR), show that PESQ and PSNR values do not always reflect MOS_A and MOS_V with high precision, thus can not provide a high accuracy for the model.

We then added PER and FR parameters in the model. It was found that adding PER and FR parameters can improve the
prediction accuracy ($R^2$) of the model. The proposed model based on PER, FR and objective MOS values is given in Eq. 7.

$$\text{MOS}_{av} = \frac{\beta_0 + \beta_1 \text{FR} + \beta_2 \ln (\text{VQ}_{\text{psnr}}) + \beta_3 \text{AQ}_{\text{PESQ}}}{(1 + \beta_4 \text{PER})^2},$$  \hspace{1cm} (7)$$

Table V shows the Pearson’s correlation coefficient ($R^2$) and Root Mean Square Error (RMSE) for both models.

<table>
<thead>
<tr>
<th>TABLE V</th>
<th>COEFFICIENTS OF METRIC MODEL</th>
</tr>
</thead>
<tbody>
<tr>
<td>Coefficient</td>
<td>Eq. 5</td>
</tr>
<tr>
<td>$\beta_0$</td>
<td>2.284</td>
</tr>
<tr>
<td>$\beta_1$</td>
<td>0.089</td>
</tr>
<tr>
<td>$\beta_2$</td>
<td>1.537</td>
</tr>
<tr>
<td>$\beta_3$</td>
<td>-</td>
</tr>
<tr>
<td>$\beta_4$</td>
<td>-</td>
</tr>
<tr>
<td>$R^2$</td>
<td>84.9%</td>
</tr>
<tr>
<td>RMSE</td>
<td>0.325</td>
</tr>
</tbody>
</table>

The scatter plots of predicted vs. subjective MOS are shown in Fig. 6(a and b).

![Fig. 6. Predicted vs. MOSav (a) Model from Eq. 5 (b) Model from Eq. 7](image)

Form Table V and Fig. 6 it can be seen that the highest accuracy (93%) is achieved for the model which combined both PESQ/PSNR and FR/PER. The model that considers FR and PER only achieved 84.9% accuracy. This suggests that the features of FR and PER are not enough to accurately predict audiovisual quality. Other parameters linked with voice/video or network (e.g., video content type, packet loss location, type of lost frames) will have an effect on audiovisual quality. This will be considered in the future work.

V. CONCLUSION

In this paper, we investigated the effects of individual modalities on end-to-end perceived audiovisual quality and analyzed the behavior of audiovisual quality for a set of selected parameters. We found that both AQ and VQ contribute to the perceived audiovisual quality and the additive model of AQ+VQ provides the highest accuracy. We also considered the effect of transmission errors encountered in wireless networks and proposed a regression-based, reference-free quality metric for the wireless H.263 and G.711 video conferencing applications. Our proposed model based on PER and FR can predict AVQ in wireless networks with an acceptable accuracy. We also studied the interaction of PSNR and PESQ objective measurements with subjective data and proposed models that can predict audiovisual quality based on these objective measures and a combination of PESQ/PSNR and network parameters (i.e., PER and FR). It was found that using PER and FR along with objective MOS values can improve the prediction accuracy of the model. This indicated the direction for future improvement of non-intrusive quality prediction model.

More subjective data are required to investigate the effects of other network and application parameters (e.g., encoders, bitrate, synchronization). Future work will be aimed at investigating more parameters to improve the accuracy of the prediction model in mobile and wireless networks.

ACKNOWLEDGMENT

The work reported here is supported in part by the EU FP7 ADAMANTIIUM project (contract No. 214751).

REFERENCES


