GRAPH-BASED PERCEPTUAL QUALITY MODEL FOR AUDIOVISUAL CONTENTS

Truong Cong Thang[1,2], *Jung Won Kang*[2], *and Yong Man Ro*[1*]
[1]Information and Communications University, Daejeon, Korea.
[2]Electronics and Telecommunications Research Institute, Daejeon, Korea.

ABSTRACT

Quality is an essential factor in multimedia communication, especially in content adaptation/compression. This paper deals with the perceptual quality of audiovisual contents. Existing audiovisual quality models, which are based on intuitive formulas to combine individual audio and video qualities, cannot clearly identify the contributions of different factors in audiovisual quality. In this paper, we present a graph-based formulation of audiovisual quality. A key advantage of our approach is that it can quantify the contributions of modalities as well as the contribution of their relation in audiovisual perceptual quality.

1. INTRODUCTION

In multimedia communication, it is well-known that quality metric is an essential factor to control compression and adaptation processes [1]. In our opinion, multimedia quality can be divided into three main categories [2]. The first category is within-modality quality, i.e. quality for a single modality (e.g. video only, audio only). The second category is cross-modality quality. For example, a surveillance video can be converted to image modality or text modality. The third category is multi-modality quality (e.g. audiovisual contents). Moreover, quality can be considered from perceptual aspect or semantic aspect [1][3][11].

As for multi-modality quality, most current research deals with audiovisual *perceptual quality*, which is also the focus of this paper. In the last decade, many studies have been carried out to investigate the factors contributing into the overall (audiovisual) quality (e.g. [4][5][6]). It is believed that the quality perceived by users is affected by various cognitive behaviors. Two important behaviors are user attention and cross-modal interaction [5]. User attention means that the user may pay more attention to one modality than the other. The phenomenon of cross-modal interaction shows that the quality of one modality is affected by the quality of another modality. That means, there exists some *coupling* (or *relation*) between video and audio modalities.

Moreover, several computational models have been proposed to estimate the overall quality from individual qualities of video and audio. Let's denote A and V the qualities of audio and video. In [6], the authors suggest that the best audiovisual quality model is an weighted sum of video and audio qualities, denoted as (V+A). Besides, different additive and multiplicative combinations such as (V·A), (V+V·A), (V+A+V·A), have been suggested in different studies (e.g. [5][7]). A possible and general explanation for this variety is the dependence of quality model on contents (or semantics).

The problem with the current models is that we can not clearly identify *the simultaneous contributions of modalities and their relation in overall quality*. For example, with the multiplicative model (V·A), we cannot identify the weight of each modality. Especially, the coupling between two modalities is not quantified in existing models.

In our previous work [3][2], the graph theory was applied to model the *semantic quality of multimedia contents*. The basic idea is that both original content and adapted content are represented by graphs, and then the similarities between the graphs are used to compute the overall semantic quality.

In this paper, we apply the graph theory to investigate the composition of audiovisual *perceptual quality*. Traditionally the graph theory seems more related to semantic aspect than perceptual aspect of multimedia contents. However, we will show that graph theory is an appropriate mathematic tool to model audiovisual perceptual quality. Specifically, a key advantage of our graph-based quality model is that it can *simultaneously quantify the weighted contributions of video and audio (due to the user attention) as well as the coupling of the two modalities.* Further, our quality model can incorporate contextual factors like user preference to model the quality of experience in multimedia communication.

This paper is organized as follows. In Section 2, we present graph-based perceptual quality model for audio-visual contents. In Section 3, subjective experiments are carried out to obtain quality models of some contents. In Section 4 we discuss the advantages of the graph-based quality model. Finally conclusion is provided in Section 5.

[*] Corresponding email: yro@icu.ac.kr

2. GRAPH-BASED QUALITY MODEL

A graph-based framework to formulate the quality of an adapted content compared to the original one was first proposed in [3], however, it is described in detail here for completeness and ease of result analysis. A content, generally called a multimedia document *MD*, is represented by *MD* = (*E*, *R*), where *E* is the set of *N* entities, and *R* is the set of relations among the entities. Each entity e_i of *E* has an attribute q_i and a weight w_i. A relation r_{ij} between entity e_i and entity e_j has an attribute v_{ij}. In terms of graph, entities are represented by nodes and relations are represented by connected edges. Here, we consider q_i as the quality of entity e_i itself ($0 \le q_i \le 1$), and v_{ij} as the strength of relation r_{ij}. Without loss of generality we let every q_i of original *MD* be 1. An adapted multimedia document is represented by $MD^* = (E^*, R^*)$, which is a graph adapted from the original graph by changing the qualities of entities, or removing some entities. We suppose that $q^*_i = 0$ if node *i* is removed.

Similar to related studies in content retrieval (e.g. [8][9]), our similarity measure is composed of *graph entity similarity* and *graph relation similarity*. The graph entity similarity is defined as the sum of match values of the entities between two graphs:

$$\begin{aligned} S_e(MD^*, MD) &= \sum_{i=1}^{N} match\,(e_i^*, e_i) \\ &= \sum_{i=1}^{N} w_i \cdot q_i^* \cdot q_i \; = \sum_{i=1}^{N} w_i \cdot q_i^* \end{aligned} \qquad (1)$$

And the graph relation similarity is defined as the sum of match values of the relations between two graphs:

$$\begin{aligned} S_r(MD^*, MD) &= \sum_{j=2}^{N}\sum_{i=1}^{j-1} match\,(r_{ij}^*, r_{ij}) \\ &= \sum_{j=2}^{N}\sum_{i=1}^{j-1} w_i \cdot q_i^* \cdot w_j \cdot q_j^* \cdot v_{ij} \end{aligned} \qquad (2)$$

By this definition, a relation will be affected when any of the two related entities has lower quality.

The quality of an adapted document *with respect to the original document* is defined by:

$$Q(MD^*) = \frac{\lambda \cdot S_e(MD^*, MD) + (1-\lambda) \cdot S_r(MD^*, MD)}{\lambda \cdot S_e(MD, MD) + (1-\lambda) \cdot S_r(MD, MD)} \qquad (3)$$

where λ ($0 \le \lambda \le 1$) is a constant that controls the proportions of entity similarity and relation similarity.

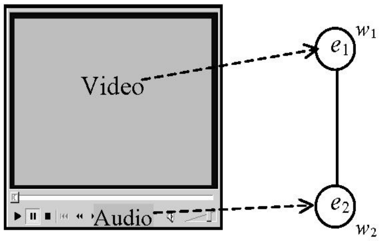


Fig. 1: Audiovisual content and its corresponding graph.

Now, we consider a practical application where an audiovisual content is streamed to the user. Here, both video channel and audio channel can be strongly scaled. For this purpose, we represent an audiovisual content by a simple graph consisting of two entities and one relation as in Fig. 1. In this graph, entities e_1 and e_2 correspond respectively to video channel and audio channel. From Eq. (3), the overall quality of an adapted audiovisual content is given by:

$$Q = \frac{\lambda \cdot (w_1 \cdot q_1^* + w_2 \cdot q_2^*) + (1-\lambda) \cdot w_1 \cdot w_2 \cdot v_{12} \cdot q_1^* \cdot q_2^*}{\lambda \cdot (w_1 + w_2) + (1-\lambda) \cdot w_1 \cdot w_2 \cdot v_{12}} \qquad (4)$$

It should be noted that, the denominators in Eq. (3) and Eq. (4) are not variable and they are used just to normalize the overall quality into the range [0, 1].

From Eq. (4), we note that one may use the value of v_{12} to control the proportions of graph entity similarity and graph relation similarity in overall quality. So we can set λ = 0.5 to simplify the model. Eq. (4) can be further rewritten as follows:

$$Q = a \cdot q_{vid} + b \cdot q_{aud} + (1 - a - b) \cdot q_{vid} \cdot q_{aud} \qquad (5)$$

where $q_{vid} = q^*_1$, $q_{aud} = q^*_2$ and *a, b* are unknown parameters with

$$a = \frac{w_1}{w_1 + w_2 + w_1 \cdot w_2 \cdot v_{12}} \quad \text{and} \quad b = \frac{w_2}{w_1 + w_2 + w_1 \cdot w_2 \cdot v_{12}} \,. \qquad (6)$$

In the above equations, the values of w_1, w_2, and v_{12}, are relative, so we can set w_1=1. And w_2 and v_{12} are obtained by:

$$w_2 = \frac{b}{a} \quad \text{and} \quad v_{12} = \frac{(1 - a - b)}{b} \,. \qquad (7)$$

From the above derivation, we can see that an audiovisual quality model (Eq. (5)) is composed of three terms: the first two are the individual *contributions of video and audio modalities*, and the third is the *contribution of the relation* between two modalities. In the next section, we will find the specific values of parameters *a* and *b* for certain contents and then compare the quality models and graph parameters of the contents.

3. EXPERIMENTS

To find quality models of audiovisual contents in streaming, we carry out subjective tests to measure qualities of audio/ video channels, and audiovisual combination. Then multiple regression is used to find the unknown parameters of Eq. (5).

3.1 Experiment settings

The test procedure is similar to that in [2][11], which is based on DCR method of [10]. Each time during the test, a subject is presented with two content versions, the original and then the adapted one, so that the subject can give a quality score to the adapted version with respect to the original one. Each score will take an integer value in Likert-style eleven-point scale, from 0 to 10. A score of 10 means that the adapted version has the same presentation quality as the original one, while a score of 0 means a very annoying presentation. The final score for each test version is the mean score of all subjects. During the real test, the test versions are shown randomly, so the subjects are not biased by a priori knowledge of presentation ordering.

Table I: Description of contents used in subjective tests.

No.	Contents	Description
1	Teacher	A teacher is speaking, similar to a head & shoulder scene.
2	Conversation	A conversation of two men, the camera is switched from one person to another.
3	Scenery	A beautiful scenery with back-ground music.

In our experiment, we select three audiovisual contents as described in Table I. The lengths of these contents are about 10s. The test contents are configured specifically for the practical audiovisual streaming service over wireless network. The video channel is coded in MPEG-4 format, with original frame rate of 30fps and frame size of 320x240. For an original audiovisual content, the video channel is adapted with four different quantization parameters (QP = 5, 15, 25, 30), while the audio channel is adapted with four different sampling rates (32KHz, 16KHz, 8KHz, and 4KHz). So, for each original content, there are 4 video versions, 4 audio versions, and 16 audiovisual versions.

The test versions are presented on a 20" Apple Cinema LCD Monitor, at the resolution of 1280x768 and with progressive display. The color of monitor background is set to 50% grey. 12 non-expert subjects were recruited to participate in the experiment.

3.2 Results and Analysis

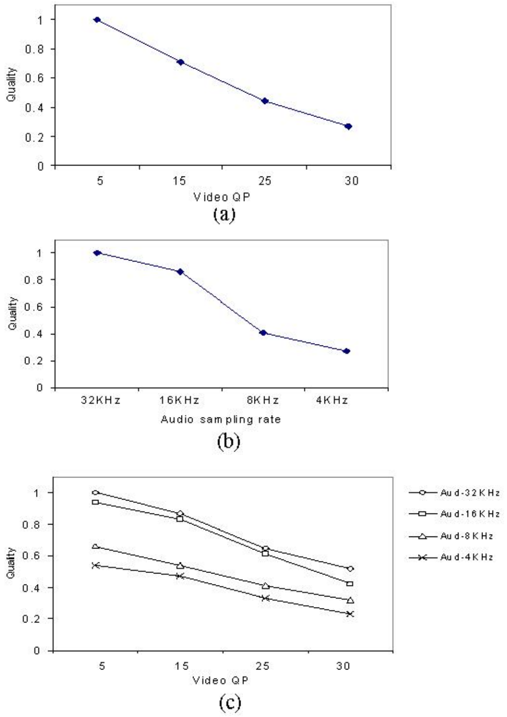
In this part, we focus on analyzing the specific audiovisual quality models. The detailed statistical analysis of subjective results will be reported in another work. The subjective scores obtained in the tests are normalized to the range [0, 1] for consistency with the above formulation. Fig. 2 shows the average scores of the audio, video, and audiovisual versions of the teacher content as a representative example. From Fig. 2 it is obvious that audiovisual quality is affected by both audio and video qualities. The impacts of video and audio channels (modalities) are rather equal for this content.

Similar to [5][6], multiple regression is applied to the subjective scores of each content to get specific quality model of that content. The parametric function to be fitted is Eq. (5). The obtained quality models for the three above contents are given as follows:

$$Q_{teacher} = 0.44 \cdot q_{vid} + 0.38 \cdot q_{aud} + 0.18 \cdot q_{vid} \cdot q_{aud} \quad (8)$$

$$Q_{conversation} = 0.32 \cdot q_{vid} + 0.43 \cdot q_{aud} + 0.26 \cdot q_{vid} \cdot q_{aud} \quad (9)$$

$$Q_{scenery} = 0.58 \cdot q_{vid} + 0.35 \cdot q_{aud} + 0.07 \cdot q_{vid} \cdot q_{aud} \quad (10)$$



(a)

(b)

(c)

Fig. 2: Subjective quality of teacher content: (a) video versions, (b) audio versions, and (c) audiovisual versions

The correlation of these models are all above 95%. From the specific quality models (8)(9)(10), the graph parameters of the contents are computed using (7) and are listed in Table II. Note that the values of graph parameters are relative and w_1 is set to 1 as mentioned above.

Table II: Graph parameters of the contents

Content	w_1	w_2	v_{12}	$w_1 \cdot w_2 \cdot v_{12}$
Teacher	1.0	0.86	0.47	0.41
Conversation	1.0	1.34	0.61	0.81
Scenery	1.0	0.61	0.19	0.12

From Table II, we can see the quantitative contribution of each modality as well as contribution of relation in a model. With teacher content, the video's weight (1.0) is just a bit higher than audio's weight (0.86), that means users pay nearly equal attentions to the two modalities (with a bit more focus on video). This is a little different from the finding of [5] where the video channel in head&shoulder content has lower weight. This may be because the video channel of teacher content has somewhat high spatial complexity. Meanwhile, with conversation content, we can see that users pay more attention to audio channel (1.34). As for scenery content, users pay much less attention to the audio channel (w_2=0.61). This is because the background music in scenery content does not convey much interesting information.

From Eq. (5), we note that the actual contribution of relation/coupling in audiovisual quality is impacted by not only the parameter v_{12} but also the weights w_1 and w_2. So the value $(w_1 \cdot w_2 \cdot v_{12})$ should be used in comparison. The relation's contribution of conversation content is the highest (0.81) and significantly higher than that of scenery content (0.12). This can be explained by the fact that the video and audio channels of conversation content have "synchronized changes" in the talk and are strongly coupled, meanwhile the two channels of scenery content are very loosely coupled.

The above experiment results show that it is possible to quantify simultaneous contributions of modalities' weights and relation in audiovisual quality. Obviously, the amounts of contributions of these factors depend on contents. The behavior of the contributing factors can be summarized as follows. When the contribution of a modality is high, that modality has more influence on overall quality. Meanwhile, when the relation's contribution is high, both modalities will have strong influence on overall quality due to the nature of the multiplicative term $(q_{vid} \cdot q_{aud})$. That is, degraded quality of any modality can reduce the value of $(q_{vid} \cdot q_{aud})$; and if both modalities are degraded, the reduction is multiplicative.

4. DISCUSSION

The above logic can be used to explain the findings and the difference in previous quality models. For instance, the multiplicative rule found in [5] is due to the fact that their test contents have strong coupling between modalities, while the missing of multiplicative term in other models [6][7] is indicative of a low coupling that may be ignored.

Note that, in previous literature, good models are selected from some intuitive formulae using correlation criterion. However, it seems that, in many cases, different combinations (e.g. V+A, A·V, V+A·V) have good and similar correlation values. One reason for this is the "rough" nature of subjective scores. Although our final quality model does not look more complex than previous formulae, it can provide insights into the behaviors of contributing factors.

Recently quality of experience (QoE) has been identified as an important goal for multimedia consumption. It is agreed that QoE should take into account contextual factors, such as user preference, usage conditions, etc. Another key advantage of our quality model is that it can incorporate contextual factors. For example, depending on contexts, the graph parameters (e.g. weights) can be modified to fit user's goal. Parameters in previous models are not functionally identified, so customization is not possible. For example, with A+A·V model type, video's weight cannot be changed.

In addition, it is possible to consider the change of relation strength in our model. For example, due to the characteristics of packet-switching networks, the audio and video channels may have timing mismatch. This can be taken into the model by modifying the strength of relation according to the degree of mismatch. On the other hand, in the multiplicative term, the qualities of video and audio may have different orders (e.g. $A^{\alpha} \cdot V^{\beta}$), so they would cause different impacts on overall quality. These issues are reserved for our future work.

We can see that an inherent difficulty of audiovisual quality is the dependence of quality models on contents. However, this problem can be overcome by classifying contents into different classes (e.g. based on their features or semantics), and then each class is represented by a common quality model.

5. CONCLUSION

Most current research on multimedia quality deals with audiovisual perceptual quality. However, existing quality models cannot identify clearly the contributions of different factors in the overall quality. This paper presented a graph-based formulation to model audiovisual quality. Subjective tests were carried out to get quality models of some audiovisual contents. The results show that our graph-based quality model can quantify the contributions of modalities as well as their relation in the overall quality. In the future, we will incorporate the contextual factors into our quality model, so as to systematically develop a QoE metric.

REFERENCES

[1] S.-F Chang, A. Vetro, "Video adaptation: concepts, technologies, and open issues", Proceedings of the IEEE, vol. 93, pp.148-158, Jan. 2005.

[2] T. C. Thang, Y. S. Kim, C. S. Kim, Y. M. Ro, "Quality Models for Audiovisual Streaming", Proc. SPIE Electronic Imaging, Vol. 6059, Jan. 2006.

[3] T. C. Thang, Y. J. Jung, Y. M. Ro, "Semantic Quality for Content-Aware Video Adaptation", Proc. IEEE MMSP2005, Shanghai, Oct. 2005.

[4] M.P. Hollier, A.N. Rimell, D.S. Hands, R.M. Voelcker, "Multi-modal perception", BT Technology Journal, Vol. 17, No. 1, pp. 35-46, Jan. 1999.

[5] D. S. Hands, "A basic multimedia quality model," IEEE Trans. Multimedia, pp. 806–816, Dec. 2004.

[6] S. Winkler, C. Faller, "Perceived Audiovisual Quality of Low-Bitrate Multimedia Content", IEEE Trans. Multimedia, Vol. 8, No. 5, pp. 973- 980, 2006.

[7] ETSI TISPAN Tech. report, "Review of material available on QoS requirements of multimedia services", 2005.

[8] S. Berretti, A. Del Bimbo, E. Vicario. "Efficient Matching and Indexing of Graph Models in Content Based Retrieval", IEEE Trans. PAMI, vol. 23, pp.1089-1105, 2001.

[9] J.-H. Lim, Q. Tian, and P. Mulhem, "Home Photo Content Modeling for Personalized Event-Based Retrieval", IEEE. Multimedia, Vol. 10, No. 4, pp. 28-37, 2003.

[10] ITU-T P.911: 'Subjective audiovisual quality assessment methods for multimedia applications', International Telecommunication Union, Geneva, Switzerland, 1998.

[11] T. C. Thang, Y. J. Jung, Y. M. Ro, "Modality conversion for QoS management in Universal Multimedia Access", IEE Proc.- Vision, Image & Signal Processing, vol. 152, Issue 03, pp.374-384, Jun. 2005.