An Audio-visual Objective Quality Model Based on BP Neutral Network

Xinlu Han[#], Yaodu Wei, Xiang Xie
School of Information and Electronics
Beijing Institute of Technology
Beijing, China
[#]hxl @bit.edu.cn

***Abstract*—This article proposes an audio-visual objective quality model, which is adaptive to the content changes. The major disadvantage of most existing models is that there is no unified model to evaluate the quality according to the content. So we firstly extracted the audio and video features and the correlation of the audio and video. Then we used BP neutral network to build the model, and used the subjective test to train and verify the model. The result of the test shows that this model is better than the second-order model.**

Keywords-audio-visual; quality evaluation; content; BP neutral network

I. Introduction

Nowadays, the development of multimedia communication service makes the communication easier and more convenient, but the quality of the multimedia service varies. The accuracy of evaluation not only can standard the service level, but also can promote the development of the industry. For the past years, the study of multimedia quality evaluation limits in single-channel, which means quality evaluation for audio or video separately. But for people, the evaluation is cross-channel, so the information extracted from audio and video is indivisible. The cross-channel evaluation is closer to the real feeling of people.

The audio-visual quality model is based on the result of audio and video's one-channel quality. The general method is suggested by D.Hands [1]. We marks video and audio quality as V and A, so the audio-visual quality Q is

$$Q = a_1 \times V + a_2 \times A + a_3 \times V \times A + a_4 \qquad (1)$$

a_1, a_2, a_3, a_4 are coefficients. To get these constants, we need to perform subjective quality evaluation, and do regression analysis on the results. Many existing results show that, the coefficients getting from different audio-visual sequences are distinguished, which means the model is only suitable for one kind of sequences. If we want to improve the results of other kind sequences, we need to do another subjective test to retrain the model, which means we don't have a fixed model [2]. Generally speaking, the main reason for the unset model is the difference of the audio-visual content [3]. In able to get a fixed model, we need to design a model to adapt to the contents.

In this paper we suggest a quality evaluation model which can adapt to the audio-visual contents. This model designs a group of content features to analyze the audio-visual contents. Finally, a neutral network combines the correlation audio/video quality and content features to get the quality of audio-visual sequences.

II. OBJECTIVE QUALITY MODEL

A. Extraction of Content Features

To describe the effect of different content to audio-visual quality, we need to analyze the weight of video and audio quality in audio-visual quality, which means analyzing the competitive relationship between audio and video [4]. When people see video and listen audio together, the evaluation of audio and video is different from the evaluation when audio and video display separately. This means expect for the competitive relationship audio and video have connection, and this is called cross-modal interaction.

Studies [5] show that the attention difference between audio and video will effect cross-modal interaction. So the competitive relationship is widely recognized. The distribution of attention is related to the radio of audio and video attention and the correlation of audio and video.

In our model, we extract three kinds of features: visual attention, auditory attention and audio-visual correlation. We specified the three kinds into six features, shown in Fig.1.

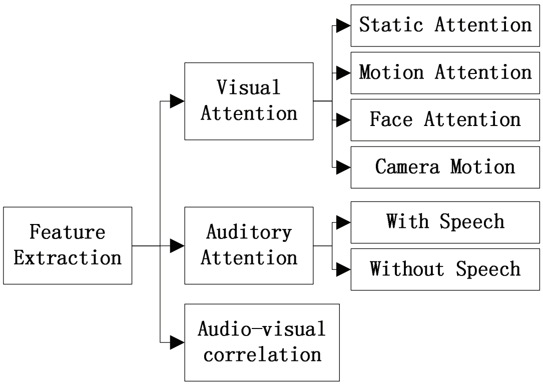

Figure1. Feature Extraction

1) Visual Attention

The State Major Science and Technology Special Projects (2010ZX03004-003)
Supported by Program for Changjiang Scholars and Innovative Research Team in University (PCSIRT)

978-1-61284-774-0/11/$26.00 ©2011 IEEE

a) Static Attention

Static attention $\vec{S}$ is calculated by algorithm proposed by Itti and so on. We can get a saliency map from each frame of images. If the map's value is higher, the saliency of this pixel is higher [6].

b) Motion Attention

We use Yufei Ma's method and consider the motion's speed and direction [7]. Firstly we need to do full search to video's motion vector, and calculate the adjacent frame's motion vector. By the picture of motion vector, we can observe the speed and direction information. The objects having significant moving display the similar motion direction. We calculate every vector's neighbor to get the motion direction, if the direction is consistent, this frame may have the significant motion, if the direction is mess, the conclusion is adverse. We calculate 25 frames before and after to get motion attention $\vec{B}$.

c) Face Attention

The area of recognizable face in the video is larger, the attention is more attractive to people. Our model detects every frame's face and uses the average area of the face as this video's face attention $\vec{F}$.

d) Camera Motion

We use Lin Liu's method to get the translation of the camera in horizontal and vertical direction and camera's zoom ratio. And then calculate the average of these three values to get camera motion $\vec{C}$.

2) Auditory Attention

When we listen to the audio, the sound with information is always attractive, otherwise, the sound without information is always considered as noise. So the judgment of auditory attention can transform to count the information of the audio. In audio, the speech has the obvious semantic and contains lots of information, other types of audio usually show emotions by music, so we divide the audio into speech and without speech.

We use HTK tool and 863 Mandarin Speech Recognition Library to do speech recognition. Sometimes the music may be mistakenly identified as speech, but music usually has longer length. So we define single syllable which has more than one second duration as a false identification. We calculate the proportion of the speech in the audio as the attention, marked as $\vec{A}$.

We use music emotion's detection mentioned in article [8]. We divide the music into vitality, calm, sad and happy and the max probability for these four kinds as the attention, marked as A .

3) Audio-visual Correlation

According to Gestalt psychology, the audio and video event, which are simultaneous and isotropic, is usually understood as the same event. When the simultaneity is destroyed, the audio and video have a closer correlation, the subjective quality decreases more obviously. Our model uses the method proposed in article [9], and artificial introduces the synchronizing errors. We calculate the correlation by using co-inertia analysis in different synchronization error, and get the correlation curve. Using Gaussian function to fit and describe the curve's convexity, and using Gaussian's variance σ^2 as the correlation's feature, marked as $\vec{R}$.

B. *Neural Network*

1) The selection of neural network

This objective model has multiple input variables, so we want to build an ideal non linear system which has the capacity of infinite approximation. There are many types of artificial neural network, and the multi-level mapping back propagation network is suitable for our model. So we choose BP neural network model.

The basic idea of BP algorithm is the learning process is composed of signal's forward propagating and error's antipropagation. During forward propagating, input sample transport from the input layer, and transport to the output layer after processing by the hidden layer. If the real output is different from the ideal output, the error needs to transport to antipropagation period. This period is transporting the error from the hidden layer to the input layer and getting each layer's error signal, which uses as the correction basis. This progress is the study and training of the network. If there is no limit of the nodes of the hidden layer, two layers (with only one hidden layer) BP network can realize arbitrary linear mapping. So we choose two layers' model.

2) The audio-visual model based on neural netwok

In this article we use MATLAB's NNbox and nntool to realize our model.

III. SUBJECTIVE TEST

A. *Test Settings*

The subjective test environment conformed to ITU-T P.911 recommendations [10].

For the equipment, four Samsung T220P LCD monitors with the original resolution of 1920 × 1080 and four Sennhesire HD25 earphones were used. Video sequences were played in front of a uniform gray background, with their original size in the center of the screen.

A nine level ACR scale was used to get MOS as recommended in ITU-T P.911. During the test, the subjects were asked to give their opinions according to their acceptance of the audio-visual quality after each sequence, and the grade is integer from 1 to 9.

The test used 48 subjects, 24 males and 24 females, aging from 20 to 30. All the subjects were reported normal or correct to normal vision and normal hearing. None of the subjects were the experts in audio or video and none of them were familiar with the sequences. Four subjects were in a group and having test together, but each group had a shuffle order.

The test was divided into three stages. The first stage, we only played video and get the one-channel video grade. The second stage, we get the one-channel audio grade. The last stage, we played audio and video together and the subjects need to grade for audio, video and audio-visual quality.

B. *Test Secquences*

We have 32 audio-visual sequences in this test, and the length for each sequence is about 5 seconds. All the test sequences are transferred from high-quality source video to 320 × 240 resolution, uncompressed avi format video. The distribution of audio-visual sequences in time intensity (TI) and space intensity (SI) shows in Fig.2, it shows that the chosen sequences can cover the common content.

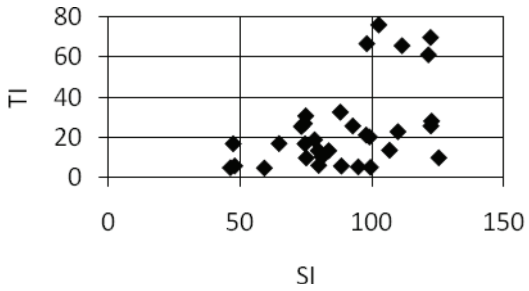

Figure2. Scatter plot of test sequences in SI-TI plan

C. *Test Result*

We use the data of the subjective test to do second-order polynomial modeling. Firstly we study the relationship between video quality (V), audio quality (A) and audio-visual quality (Q). After rotating the second-order curve surface, we get Fig.3. It shows V , A , Q can make a thick curve surface, which may due to the content's effect and we cannot model it.

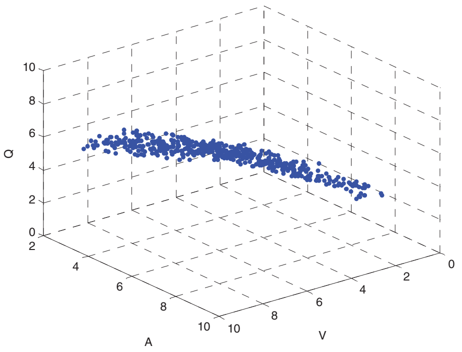

Figure3. Relationship between V, A and Q after rotation

By using the same method, we study the video quality with audio (V_{cross}), the audio quality with video (A_{cross}) and the audio-visual quality (Q). We also rotate the second-order curve surface and get Fig.4. This time we get a thin surface, so we can model it. And the result is

$$\begin{aligned} Q = &0.0101469 + 0.570246 \times V_{cross} + 0.365787 \times A_{cross} \\ &+0.0692957 \times V_{cross} \times A_{cross} - 0.037644 \times V_{cross}{}^{2} \\ &-0.026644 \times A_{cross}{}^{2} \end{aligned} \qquad (2)$$

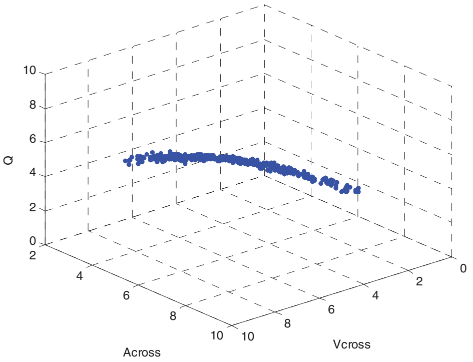

Figure4. Relationship between Vcross, Across and Q after rotation

IV. MODEL PERFORMANCE

The artificial neural network gets the model from training data, and then verifies the model by testing the data. So in our work, we divided our data into two groups. One group of data uses for training, the other group of data uses for testing. In this model, we used 10 sequences as the test sequences. The other 22 sequences' subjective test results were used as the training data.

In the design process, the structure of BP neural network determines the simulation results of the model somehow. BP network's input and output nodes are determined by the nature of the practical problem. In our model, we have 8 inputs and 1 output. The most critical factor that affects the network structure is the number of neurons L in hidden layer. There is no uniform standard to calculate L, so we often determine it according to empirical formula. Generally, we use three ways to select the best number of neurons in hidden layer. In the following, the number of input and output neurons is marked as M and N [11].

- $\sum_{i=0}^{M} C_{M_i}^{i} > k$, k is the number of samples, M_i is the number of units in hidden layer, i is the constant in $[0, M]$.
- $L = \sqrt{M + N} + a$, $a$is the constant in $[1,10]$.
- $L = \log_2 M$.

According to empirical formula and verifying the best number of unites in hidden layer, we chose 5 neurons in hidden layer in our model.

In the audio-visual objective quality evaluation model based on neural network, we extracted 6 features from audio and video content, and we got audio quality and video quality from subjective test. These 8 features are the input of the model, which means the multi-dimensional random variable X .

$$X = \left[V\ A\ \vec{S}\ \vec{B}\ \vec{F}\ \vec{C}\ \vec{A}\ \vec{R} \right]^{T} \qquad (3)$$

We use MSE (Mean Square Error) to quantitatively describe the model's performance. The smaller the MSE is, the model's result is closer to the subjective test's result.

We defined the error of each measured value as $\varepsilon_1\varepsilon_2\cdots\varepsilon_n$, and MSE is:

$$MSE = \sqrt{\frac{\varepsilon_1^2+\varepsilon_2^2+\cdots+\varepsilon_n^2}{n}} = \sqrt{\frac{\sum\varepsilon_i^2}{n}} \tag{4}$$

The BP neural network's MSE1=0.1057. Fig.5 shows the scattergraph of subjective test and the prediction result based on BP neural network.

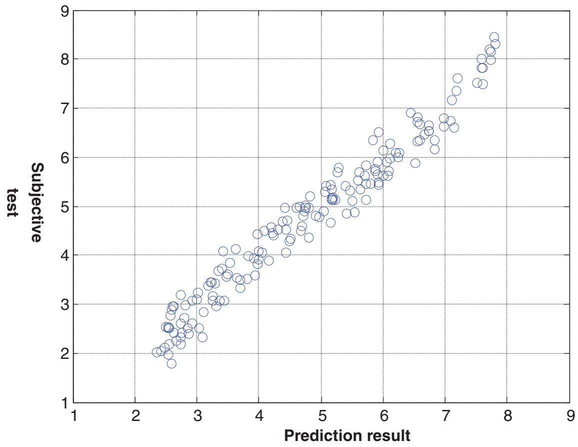

Figure5. the scattergraph of BP neural network

We also calculate the second-order model's MSE. The second-order modal cannot adjust to all the sequences, so in this paper, we only put one of the sequences' scattergraph in Fig.6. We calculated all the MSE of this model, and had the average value MSE2=0.1362.

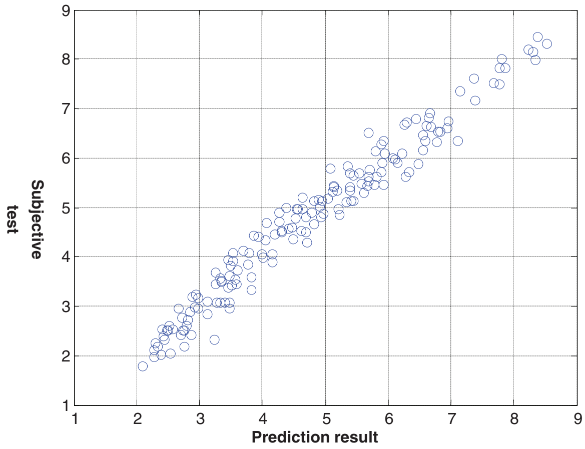

Figure6. the scattergraph of second-order model

V. CONCLUSION

In this paper, we built an audio-visual objective quality model based on BP neural network. In this model, we not only concluded the content of the multimedia, but also considered the cross-relation in audio and video. So our model is a universal model.

We compared the BP neural network model and second-order model's performance. The MSE of BP model is smaller than that of second-order model. The second-order model is not fixed and it changes with different sequences. We need to build the model for each sequence, and each model has a MSE. We calculated the average of all these MSEs marked as this model's performance. Sometime we may not find the suitable second-order model, so the MSE is higher and made the result higher.

The BP neural network model is suitable for all the sequences, because it considers the content. We have 8 input variables, which consider nearly all the type of audio and video content, so it has a better performance and lower MSE.

REFERENCE

[1] Hands, D.S., "A basic multimedia quality model," IEEE Transactions on Multimedia,2004 ,12 (6): 806-816.

[2] GOUDARZI M, LINGFENG S, IFEACHOR E, Audiovisual Quality Estimation for Video Calls in Wireless Applications, GLOBECOM 2010, Miami, U.S., Jan. 10,2010,1-4.

[3] KHAN A, LINGFENG S, IFEACHOR E, Impact of video content on video quality for video over wireless networks, Fifth International Conference on Autonomic and Autonomous Systems, 2009, Valencia, Spain, May 26, 2009, 277-282.

[4] NAVARRA J, ALSIUS A, FARACO S S, SPENCE C, Assessing the role of attention in the audiovisual integration of speech, Information Fusion, 2010, 11(1): 4-11.

[5] Rimell A ,Owen A. The effect of focused attention on audio-visual quality perception with applications in multi-modal codec design. In: Proceedings of 2000 IEEE International Conference on Acoustics, Speech, and Signal Processing. Istanbul, 2000, 6 : 2377.

[6] ITTI L, KOCH O, NIEBUR E. "A model of saliency based visual attention for rapid scene analysis," IEEE Transactions on pattern analysis and machine intelligence. 1998, 20(11):1254-1259.

[7] YUFEI M, XIANSHENG H, LU L, HONGJIAN Z. "A generic framework of user attention model and its application in video summarization," IEEE Transactions on Multimedia. 2005, 7(5):907-919.

[8] ZHANG F X. Classification of emotional music retrieval based on system design and implementation. Beijing, Beijing Institute of Technology, 2007.

[9] WEI Y D, XIE X, KUANG J M, HAN X L. A speech-video synchrony quality metric using CoIA, 18th International Packet Video Workshop(PV), 2010. HongKong, Dec. 13-14, 2010, 173-177.

[10] ITU-T Recommendation P.911, "Subjective audiovisual quality assessment methods for multimedia applications," 1998.

Quan H. Thu, Ghanbari M. "A comparison of subjective video quality assessment methods for low-bit rate and low-resolution video," 7th IASTED, pp. 70-76. 2005.

[11] TIAN G Y, HUANG H Y. the determination of hidden layer in neural network, Information Technology, 2010(10): 79-81.