Evaluation on Perceptual Audiovisual Delay using Average Talkspurts and Delay

Jian Wang, Fuzheng Yang, Zhiqing Xie
State Key Lab. of Integrated Service Networks
Xidian University,
Xian, China

Shuai Wan
School of Electronics and Information
Northwestern Polytechnical University
Xian, China

***Abstract*—In this paper, six different conversation tasks with pre-determined content, referred as 'non-free conversations', are designed for evaluation of audiovisual delay. Extensive subjective experiments were performed to evaluate the delay perception. Experimental results show that subjective audiovisual quality is getting worse with the increase of delay, while delay perception is closely linked to the characteristics of audiovisual conversations. Using talkspurts to capture the temporal characteristics of non-free conversations, a logarithmic function or quadratic polynomial may be appropriate to model the perception delay.**

Keywords-delay, audiovisual quality, non-free conversation tasks, mean opinion score

I. INTRODUCTION

Benefited from the rapid development of broadband IP and multimedia technologies, the related applications have gained a steady increase lately. Various multimedia applications prevail and are becoming commonplace over the Internet, such as Voice over IP, videoconferencing, and videophones. Since quality generally is not guaranteed in an IP network, transmission delay has a great impact on users' quality experience in real-time audiovisual communications over internet. Therefore, it is of great importance to evaluate the delay effects which users perceive. Some researchers in NTT proposed a linear model [1] to integrate audiovisual delay and the mean opinion score(MOS), where a task referred to as name guessing in ITU-T Rec. P.920 [2] was adopted. A more interactive task named LEGO building blocks model was proposed in [3], where the results showed that subjects can tolerate delay to a large extent and the subjective quality is not linearly decreasing with delay, but vibratory. Considering conversational tasks, however, it is well-known that the quality is significantly dependent on the specified task performed in subjective quality assessment [4]. From the users' viewpoint, it is natural that the perceived quality varies according to how crucial or demanding the conversation task is. So the derived conclusions based on the mono tasks above are limited to typical applications and are apparently not universal. In order to find a general conclusion which applies to different applications, it is necessary to extract common parameters from different conversations which dominantly influence the users' delay perception. Many researchers designed experiments to empirically evaluate how different parameters influence delay perception in speech conversations. One typical example is that N. Kitawaki demonstrated that delay perception is greatly influenced by temporal characteristics of conversational speech, such as talkspurts of a conversation, conversation switching times from one side to the other, deviation of signal duration and utterance speed [4]. On the other hand, a temperature metric for conversation interactivity based on the four states conversation model was proposed by F. Hammer, where the four states are defined as talkspurts of either two sides, double talk and mutual silence, respectively [5].

Although the above models measure delay perception through temporal parameters, delay in itself is not listed in the candidates. In multimedia quality assessment, however, delay is usually available as a parameter provided by network inspection or primary analysis of the bit-stream. It is therefore more straightforward and reasonable to employ the delay time as a variable of delay perception. As a substantial distinction from existing work, we propose to incorporate delay in evaluation of delay perception.

Through extensive experiments and related analyses, we conclude that delay perception can be primarily modeled using delay and one another temporal parameter, namely, average talkspurts of sentences in a conversation. Although more parameters, such as utterance speed and deviation of signal duration, can better represent a conversation, the corresponding experiments are much more difficult to handle, resulting in complicated data processing and lack of confidence of related data. On the other hand, the slower the utterance speed is, the longer the talkspurts would be, considering the same sentence. That is, the utterance speed can be well captured by the talkspurts. Similarly, the deviation of signal duration can be derived from talkspurts since it is in essential the same as the variation of the talkspurts between different sentences. Therefore, talkspurts are actually very effective to represent temporal characteristics of a conversational speech. Moreover, talkspurts are very easy to extract from conversations. Accordingly, we will study the relations among average talkspurts of different sentences, audiovisual delay and mean opinion score in this paper. The corresponding observations and conclusions are essential in design of objective quality assessment metrics for multimedia services where audiovisual conversations suffer from network delay.

The rest of this paper is organized as follows. Section II provides details of subjective experiments on perceptual audiovisual delay in videophone communications, where the

978-1-4244-6516-3/10/$26.00 ©2010 IEEE

experiments settings and designed conversation tasks are specified respectively. Corresponding experimental results are presented and discussed in section III. This paper closes with conclusions given in section IV.

II. SUBJECTIVE EXPERIMENTS ON PERCEPTUAL AUDIOVISUAL DELAY

For the evaluation of perceptual audiovisual delay considering different average talkspurts, the videophone scenario was employed in all experiments, as a typical example of interactive video applications, where the videophone software is developed by our research group following ITU-T H.323 [6]. In this scenario, several conversation tasks different in average talkspurts were addressed. Details about the experiments are given below.

A. Test Configurations

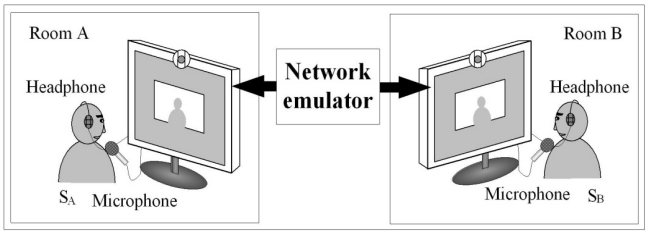

Figure 1. Subjective Delay Test System

Figure 1 depicts the experimental setup used for the experiments, where two identical videophone terminals were placed in two separated rooms for conversational purposes. The microphones and headphones are of excellent quality. Several different delay periods, ranging from 100msec to 1400msec, were randomly introduced to the audiovisual streams for each task in a network emulator for simulations. Other detailed settings and configurations of the experiments are listed in Table I.

TABLE I. FORMATS AND SETTINGS

Audio coder	AMR
Video coder	MPEG4
Video format	QCIF
Delay[ms]	100, 250 ,450,600,800,1000,1400

Notice:The minimum end-to-end delay was not 0 ms but about 100 ms, which was mainly caused by video encoding and decoding procedures.

B. Test scenarios

Although free conversations are what actually happen in practice, it is not realistic to use free conversations in the experiments for delay evaluation. This is mainly because the sentences in a free conversation vary much in length. Moreover, free conversations will incur lots of double talks and interruptions due to existence of delay [5][7][8], which brings much inconvenience to figure up the related talkspurts. Correspondingly, non-free conversations are employed in this paper for the evaluation of delay effects using talkspurts. The issue of using free conversations is currently under investigation.

Considering speech delay in conversations, ITU-T Rec. P.920 has listed some standard non-free conversation tasks for evaluation. However, we address the audiovisual delay in multimedia applications where the video part cannot be ignored under perception. The video sometimes looks like freezed during the long delays as the peer stares at the screen still while waiting for responses, which helps identifying delay. Moreover, those tasks are not appropriate for Chinese subjects to perform. Accordingly, the tasks defined in ITU-T Rec. P.920 cannot be directly used in our experiments. Based on ITU-T Rec. P.920 and following its spirit, as well as combined with the speaking way of Chinese, six new non-free audiovisual conversation tasks are designed and used in our experiments. The involved tasks are listed in Table II. The employed tasks vary in average talkspurts and are more interactive. Subjects can perform these tasks without referring to materials, and the delay effect can be evaluated from the speech conversation as well as from the video through their partner's facial expressions. In each task, there are at least five interactions, so subjects in both sides have several chances to perceive delay. Each sentence in a given task is of the same length.

The six tasks listed in Table II are shortened as "count", "week", "poem", "long poem", "parallelogram" and "fruit" respectively for conciseness in the following discussions. Average talkspurts of these tasks are 0.3 second, 0.6 second, 1.1 seconds, 3.8 seconds, 6 seconds and 10 seconds respectively.

TABLE II. AUDIOVISUAL CONVERSATION TASKS

Tasks	Details
count	Take turns in counting numbers from one to ten
week	Take turns in speeking weeks from Monday to Sunday
poem	Take turns in speaking seven-character poem "The Inlaid Harp"
long poem	Take turns in speaking chinese poem "Prelude to Water Melody"
parallelogram	Take turns in speaking parallelogram theorems
fruit	Take turns in describing different fruits with long sentences

C. Test procedure

Series of experiments have been launched following the above directions. Because there are six tasks in the experiments, where each of the tasks has 7 different delays, so there are 70 test items in all. All test items are randomized in five different orders before they are presented to pairs of subjects for the purpose of eliminating order effect.

There were 24 subjects involved in the experiments, including 12 males and 12 females. Before the actual assessment, the handling of the experimental system is explained to the subjects, where they have a trial phase of two minutes to get familiar with the system. The subjects were informed to specifically evaluate the delay effect. For that purpose, the quality of the audio and video was constantly good

with neither compression distortions nor additional transmission errors introduced. By such a design, the subjects focus on the delay effect in conversations and the evaluated quality is not influenced by the other factors except for delay.

In all experiments, The single stimulus continuous quality evaluation (SSCQE) [9] is used in the subjective test. Subjects talked in Chinese during the whole conversations. They evaluated the delay quality in real time using a slider device and a continuous grading scale marked with "Excellent", "Good", "Fair", "Poor" and "Bad" respectively. The obtained mean opinion score (MOS) were collected during experiments. All audiovisual conversations were recorded for post-analysis purposes. After the experiments, all recorded sequences were processed to extract the talkspurts of each sentence. For each task, the MOSs and talkspurts were averaged over all subjects. Detailed experimental results are given below.

III. EXPERIMENTAL RESULTS AND EVALUATIONS

In this section, we present the results of all recorded conversations. Results are illustrated from two different aspects. Detailed experimental results are given below.

A. The Effect of Delay

The relation among delay, talkspurts and MOS is illustrated in Fig. 2, where the MOSs decrease monotonously with an increase of delay regardless of the conversation tasks. All the curves start from a similar level in Fig. 2 when delay is very small. For example, at the delay of 100ms, the perceptual quality of "count" is slightly smaller than those of all the other tasks, which are almost of the same high score, i.e. 4.2. Although different in scores, the perceptual quality of "count" still remains in "Excellent". The above observations show that in terms of subjective quality rating, the subjects cannot feel any delay at such small delay for any task. However, when delay is increasing, the differences between tasks become more and more obvious, where different tasks show distinctly different declining trends. Specifically, the tasks with longer talkspurts decline in a comparatively slower manner. When delay increases to 1400ms, the corresponding subjective quality for "fruit" drops to "Fair", while for "count" drops to "Bad" in the continuous grading scale.

Of all the conversation tasks, the "count" task, with the minimum talkspurts, is most sensitive to delay. It is noticeable that the MOSs obtained by our experiments are different from [10][11][12] for a similar "count" task, where his observations show that when delay is larger than 180ms, MOSs dropped to below 1. The rapid drop of MOSs in those experiments is due to the use of double-stimulus continuous quality-scale [9] method in his experiments. Apparently, delay is much easier to be perceived when a reference of no delay presents.

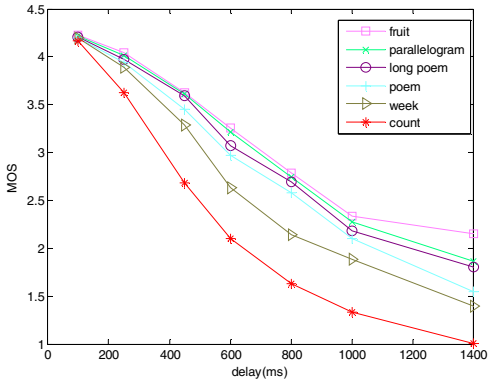

Figure 2. Relation between delay and MOS

B. The Effect of talkspurts

The relation between talkspurts and the MOS under different delay periods is depicted in Fig. 3, where the MOSs increase as a logarithmic function or quadratic polynomial with the average talkspurts of sentences. It is very clear that when delay increases, the MOSs drop correspondingly. Furthermore, with the same delay, MOSs increase rapidly with the average talkspurts, when the average talkspurts are within 2 seconds; however, when the average talkspurts are larger than that, the corresponding subjective qualities increase much more slowly. For example, at the same delay of 450ms, with a difference of 0.3 second of the average talkspurts between "count" and "week", the corresponding difference between MOSs is about 0.6, whereas for "parallelogram" and "fruit", the corresponding difference between MOSs is no larger than 0.02 point.

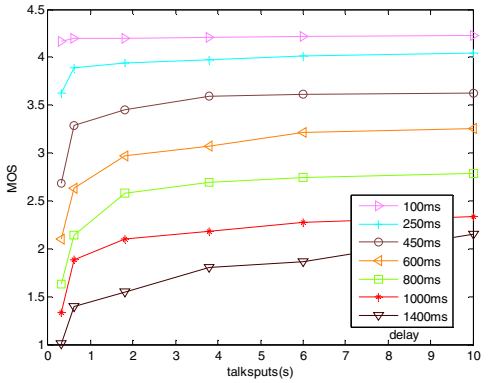

Figure 3. Relation between average talkspurts and MOS

IV. CONCLUSIONS

In this paper, we evaluated the delay perception on different delays and average talkspurts of sentences. The experimental results show that MOSs decrease with an increase of delay regardless of the conversation tasks. More importantly, the observations imply that talkspurts can well represent temporal characteristics of non-free conversations, and the longer the talkspurts of a sentence, the harder for the subject to perceive delay, where a logarithmic function or quadratic polynomial may be appropriate to model the relation. Future work includes

an extension of this method to free conversation scenarios, and to a universal objective quality assessment metric for delay.

ACKNOWLEDGEMENT

This work was supported by the National Science Foundation of China (60902052, 60902081), the Doctoral Fund of Ministry of Education of China (No. 20096102120032) and the 111 Project (B08038). The supports provided by HuaWei Technologies Co. Ltd. for subjective quality assessment are also gratefully acknowledged. Moreover, the authors would like to give a special thanks to all the subjects taken part in the tests.

REFERENCES

[1] T. Hayashi, K. Yamagishi, T. Tominaga, and A. Takahashi, "Multimedia Quality Integration Function for Videophone Services," Proc. IEEE Global Telecommunications Conference, pp. 2735-2739, Nov. 2007.

[2] ITU-T Recommendation P.920, "Interactive Test Method for Audiovisual Communications," Geneva, May 2000.

[3] F. Brauer, M.S. Ehsan, and G. Kubin, "Subjective Evaluation of Conversational Multimedia Quality in IP Networks," Proc. IEEE 10th Workshop on Multimedia Signal Processing, pp. 872-876, Oct. 2008.

[4] N. Kitawaki, and K. Itoh, "Pure Delay Effects on Speech Quality in Telecommunications," IEEE Journal on Selected Areas in Communications, vol.9, no.4, pp. 586-593, May 1991.

[5] F. hammer, P. Reichl, and A. Raake, "The Well-Tempered Conversation: Conversation Interactivity, Delay and Perceptual VoIP Quality," Proc. IEEE ICC'05, Seoul, Korea, May 2005.

[6] ITU-T Recommendation H.323, "Packet-Based Multimedia Communications Systems,",Geneva, Switzerland, Feb. 1998.

[7] F. hammer, P. Reichl, and A. Raake, "Elements of Interactivity in Telephone Conversations," Proc. in 8th International Conference on Spoken Language Processing (ICSLP/INTERSPEECH 2004), Jeju Island, Korea, Oct. 2004.

[8] F. Hammer, and P. Reichl, "How to Measure Interactivity in Telecommunications," 44th FITCE Congress 2005, Vienna, Austria, Sept. 2005.

[9] ITU-R Recommendation BT.500, "Methodology for Subjective Assessment of the Quality of Television Pictures," Recommendations of the ITU, Radio Communication Sector, Geneva, 2002.

[10] T. Kurita, S. Lai, and N. Kitawaki, "Effects of Transmission Delay in Audiovisual Communication," IEICE Trans. Commun., vol. J76-B-I, no. 4, pp. 331-339, April 1993.

[11] T. Kurita, S. lal, and N. Kitawaki, "Assessing the Effects of Transmission Delay-Interaction of Speech and Video," 14th Int. Symposium on Human Factors in Telecommunications, HFT'93, pp. 111-120, Darmstadt, Germany, May 1993.

[12] S. lai, T. Kutita, and N. Kitawaki, "Quality Requirements for Multimedia Communication Services and Terminals-Interaction of Speech and Video Delays," Proc. IEEE Global Telecommunication Conference, pp. 394-398, Houston, USA, Nov. 1993.