# Impact of Audio on Subjective Assessment of Video Quality in Videoconferencing Applications

Michael R. Frater, John F. Arnold, and Abedin Vahedian

*Abstract*—**In the real world, we commonly receive information simultaneously through two or more senses, with the brain fusing this data to produce a single coherent message. Lip-reading is one example of this phenomenon. Laboratory studies, on the other hand, often measure the response to a stimulus by a single sense and extrapolate these results to predict real-world behavior. In this paper, we show that semantics have a significant impact on viewers' sensitivity to the quality of a video sequence for spatially separated parts of the sequence and, more importantly, that this difference in sensitivity can be changed by the presence of an audio signal. This result is important for any testing of subjects' responses to visual material. One example is the subjective assessment of the quality of video in an audio–visual communications system (such as television or videoconferencing).**

*Index Terms*—**Joint audio–visual coding, joint audio–visual subjective testing, video coding, video subjective quality, video subjective testing, videoconferencing.**

## I. Introduction

In the real world, we commonly receive information simultaneously through two or more senses, with the brain fusing this data to produce a single coherent message. Lip-reading is one example of this phenomenon [1]–[4]. Subjective assessment of visual test material by human viewers is commonly used in fields where no satisfactory means of objective assessment is available. Applications include marketing surveys and the assessment of video quality for television, where results obtained strongly influence the design of equipment. Much work has been carried out to define test conditions under which reliable assessments can be performed, e.g., methods for subjective assessment of television picture quality are defined in [5], [6]. These definitions, however, are based on the assumption that valid results can be obtained by evaluating video without accompanying audio.

Much material used for subjective evaluation contains some parts that are semantically more important than others. These important regions are often referred to as the "foreground". In a video conference, for example, the head and shoulders of a speaker are more important than the background behind the speaker. This paper reports the results of experiments that show that:

1) picture degradation in the foreground has a greater impact on subjective picture quality than degradation in the background;
2) the difference between sensitivity to foreground and background degradation is increased by the presence of audio corresponding to speech of the foreground person;
3) where both foreground and background are degraded equally, there is no significant difference between the perceived degradation with and without audio.

Manuscript received March 15, 1999; revised May 18, 2001. This paper was recommended by Associate Editor T. Chen.

M. R. Frater and J. F. Arnold are with the School of Electrical Engineering, The University of New South Wales, Australian Defence Force Academy, Canberra 2600, Australia (e-mail: m.frater@adfa.edu.au; j.arnold@adfa.edu.au).

A. Vahedian was with the School of Electrical Engineering, University College, The University of New South Wales, Australian Defence Force Academy, Canberra 2600, Australia. He is now with I.R.O.S.T., Mashad, Iran (e-mail: vahedian@istn.irost.com).

Publisher Item Identifier S 1051-8215(01)08094-6.

## II. Experimental Method

The test was performed using the double-stimulus continuous quality scale (DSCQS) assessment [4] using the test conditions of [6]. In this method, a number of tests are performed serially. In each test, the viewer is asked to rate the quality of two video sequences known as "A" and "B" on a continuous scale ranging between "Excellent" and "Bad." Either A or B (chosen at random) was an original sequence with no degradation. The other sequence was degraded by the addition of white Gaussian noise. The difference between the scores for A and B is, therefore, a measure of the degradation due to the noise. All analysis is based on this difference and not on viewers' absolute quality ratings.

Testing was carried out using 49 nonexpert subjects. The complete test was viewed twice by each person. One viewing was with audio present, the other without. Participants were instructed to rate the global visual quality. They were given no instructions that one part of the image was more important than another. They were not told that one of A and B was an original sequence, nor about the nature of any degradation they might expect to observe. They were not asked to evaluate the audio quality.

In most applications for subjective assessment of video quality, sequences A and B are identical except for some form of processing. In our experiments, the background is identical. The foreground audio and video is different for each sequence so that there is no training effect due to the audio being known when sequence B is viewed. In each case, however, the same speaker is used, the location of the head and shoulders in the

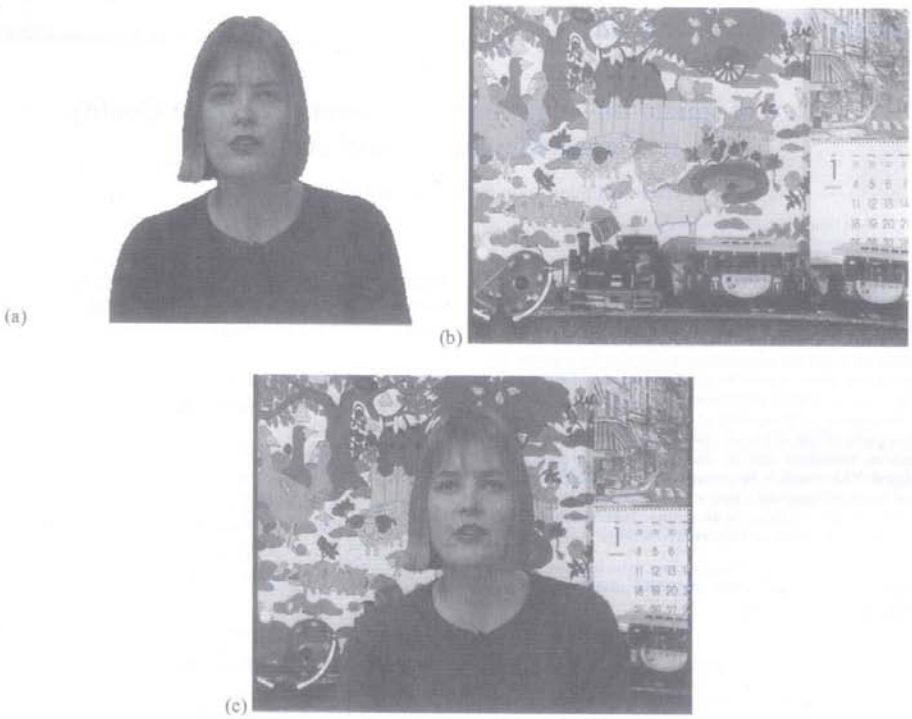


Fig. 1. First frames of: (a) foreground; (b) background; and (c) composited sequence.

video frame is essentially the same and there is little movement apart from the lips of the speaker. As a result, when viewed without audio, it is difficult to detect that sequences A and B are not identical, even knowing that this is the case.

Each video sequence was 5-s long and composited from two independent sequences. The foreground consists of a head and shoulders image of a person speaking. The background was the first 5 s of the video test sequence known as "Mobile and Calendar." The first frame of the foreground, background, and combined sequence is shown in Fig. 1.

Three types of degraded sequences were used. In the first, only the foreground was degraded. In the second, only the background was degraded. In the third, the whole sequence was degraded. Three levels of degradation were used, with noise variances of 16, 40, and 100 corresponding to peak signal-to-noise ratios (PSNR) of 36, 32 and 28 dB, respectively. The lowest noise level was chosen so that the degradation was just visible to an observer. Fig. 2 shows examples of the three types of degradation with PSNR equal to 28 dB.

## III. Results

Fig. 3 shows the results. Results are shown for each of the three different levels of picture quality (36-, 32-, and 28-dB PSNR) and for noise added to the whole picture ("ALL"), the background only and to the foreground object. These results show that:

1) viewers are more sensitive to picture degradation in the foreground than the background (with $p < 0.005$ using the Student-$t$ test);
2) the difference between sensitivity to foreground and background degradation is increased by the presence of audio corresponding to speech of the foreground person (with $p < 0.005$ using the Student-$t$ test);
3) where both foreground and background are degraded equally, there is no significant difference (with $p > 0.3$ using the Student-$t$ test) between the perceived degradation with and without audio.

The difference in subjective degradation between foreground and background for sequences with the same noise power added could simply reflect properties of the human visual system (HVS). The fact that the sensitivity to degradation in the foreground is increased by the presence of audio suggests, however, that this difference is at least partly due to the different meaning associated with foreground and background. This claim is strengthened by the very small difference between the audio/no-audio cases in subjective assessments where noise was added to both foreground and background.

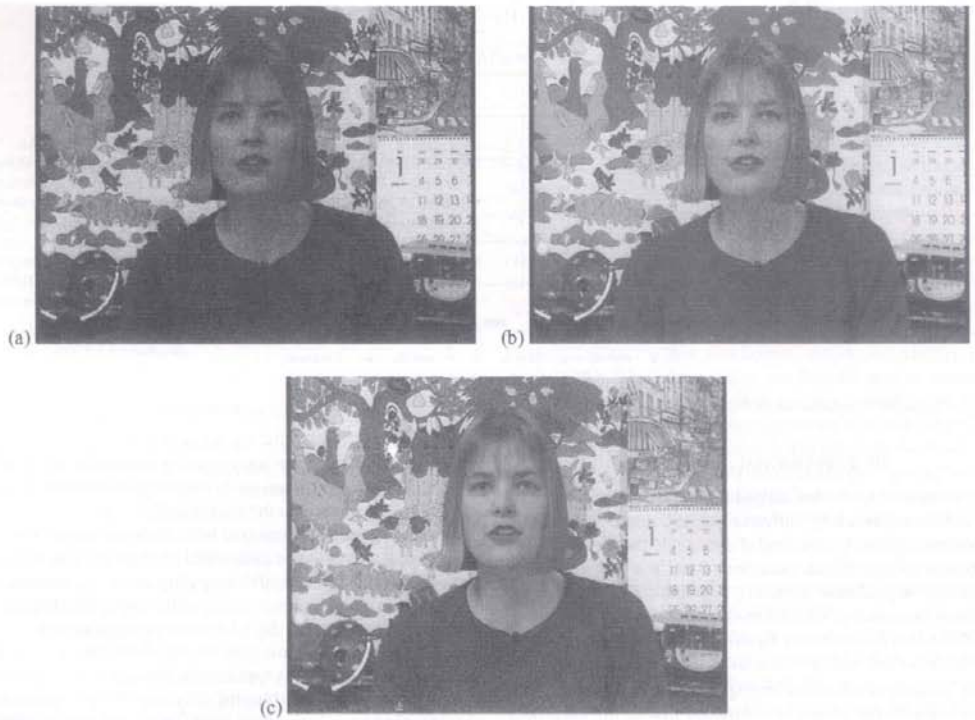


Fig. 2. First frames of sequences with degradation applied to the: (a) whole sequence; (b) foreground only; and (c) background only.

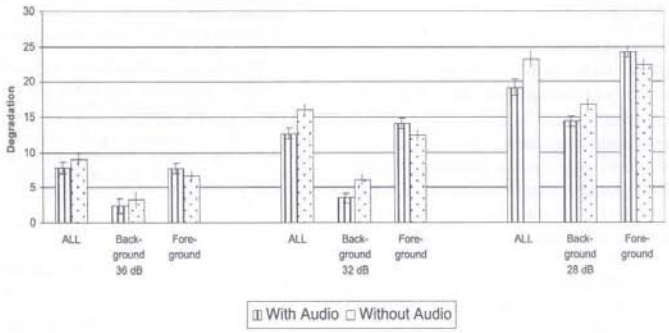


Fig. 3. Results from comparison of subjective quality with and without accompanying audio.

The change in sensitivity to degradation in the foreground and background when audio is introduced is illustrated in Fig. 4. The inner two curves are the cases without audio; the outer two are with audio. For a given subjective level of degradation, the presence of audio more than doubles the difference in RMS error between foreground and background. For a given subjective level of degradation, the presence of audio more than doubles the difference in RMS error between foreground and background. This change in perceived quality caused by the presence of audio is not directly due to properties of the HVS but to the semantics of the audio–visual scene.

The main implication of these results is that, at least in some circumstances, video subjective testing carried out with audio removed gives rise to misleading results.

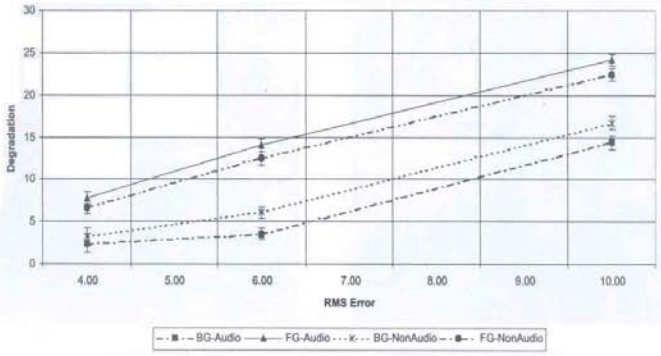


Fig. 4. Subjective degradation versus PSNR.

## IV. Experimental Verification

A number of tests were carried out to ensure that the results obtained are caused by different perceptions of degradation in foreground and background. For each of the three types of degradation (foreground, background and whole picture), the subjective degradation increases monotonically as the noise variance increases ($p < 0.005$ using the Student-$t$ test.)

Effects that do not have a significant impact, as measured by the Student-$t$ test with $p > 0.2$, include:

1) location of viewers with respect to the video screen;
2) whether the viewer was exposed first to the sequences with audio (10 subjects) or those without (39 subjects);
3) the session in which the testing was carried out.

## V. Conclusion

The results reported here demonstrate that the presence of audio has a significant impact on the subjective quality of video. These results could be usefully extended in the following ways.

- Many videoconferencing scenes contain more than one person. Our current results suggest that the viewers will be more sensitive to the image quality associated with a speaker than the background. Further work is required to establish whether the image quality associated with other people (who are not currently speaking) in the scene is also more important than the background.
- In the experiments reported here, high-quality audio was used. It remains to be established whether the same results would be obtained with low-quality audio, such as might occur in noisy environments or after coding for transmission over a low rate digital-communications system.
- The degradation introduced into the video here was white gaussian noise. Many applications introduce other types of degradation, such as blurring effects in low-resolution images and blocking artifacts due to image and video coding.

## References

[1] D. W. Massaro, *Speech Perception by Ear and by Eye*. Hillsdale, NJ: Erlbaum, 1987.
[2] H. McGurk and J. MacDonald, "Hearing lips and seeing voices," *Nature*, vol. 264, pp. 746–748, Dec. 1976.
[3] D. W. Massaro, *Perceiving Talking Faces: From Speech Perception to a Behavioral Principle*. Cambridge, MA: MIT Press, 1998.
[4] J. Driver, "Enhancement of selective listening by illusory mislocation of speech sounds due to lip-reading," *Nature*, vol. 381, pp. 66–68, 1996.
[5] J. Allnatt, *Transmitted-Picture Assessment*. New York: Wiley, 1983.
[6] "Methodology for the Subjective Assessment of the Quality of Television Pictures," ITU-R Recommendation BT.500-10, ITU, Geneva, Switzerland, 2000.