Utility Curves: Mean Opinion Scores Considered Biased

Hendrik Knoche
Computer Science Dept.
Hamburg University
knoche@tkrn.informatik.uni-hamburg.de

Hermann G. De Meer *
Dept. of Electrical Engineering
Columbia University
hdm@comet.columbia.edu

David Kirsh
Cognitive Science Dept.
Univ. of California San Diego
kirsh@cogsci.ucsd.edu

1 Introduction

Mechanisms for QoS provisioning in communication networks range from flow-based resource reservation schemes, providing QoS guarantees, through QoS differentiation based on reservation aggregation techniques to adaptation of applications, compensating for incomplete reservations. Scalable, aggregation-based reservations can also be combined with adaptations for a more flexible and robust overall QoS provisioning. Adaptation is particularly important in wireless networks, where reservations schemes are more difficult to realize. It is widely accepted that usability of Cellular or Mobile IP can be largely improved if adaptation is incorporated. Adaptation, in turn, depends on coding techniques involved and on user requirements. Those impose certain preferences and ordering relations on QoS hierarchies or degradation paths to be followed along when adaptation is performed. With the advent of MPEG4 coding techniques, more sophisticated adaptation schemes are more likely to be supported in the future. The problem still remains as how adaptation can be effectively controlled. Utility curves (UCs) have been suggested as such a mechanism. But there is a notable lack of knowledge on how utility curves can effectively be determined in a way sensibly reflecting users' needs. UCs provide a formal technique to directly relate network state, such as available bandwidth, to end-user perceived QoS. The relation is established by tests incorporating questionnaires to find out about users' opinions on certain aspects of the media's qualities presented. The quantitative result of such an assessment is called a *Mean Opinion Score*, MOS for short. An example would be a subjective MOS about when a skew between different media like audio and video becomes visible and notable to some subjects.

We feel that there are some inherent problems with this approach that might prove prohibitive for its more general application and validity. In this paper, we provide evidence for such a conclusion and outline an approach, based on more objective *Task-Oriented Performance Measures* (TPMs), which we believe is more general and suitable. This paper is a short summary of the interdisciplinary *Coqos* project, jointly pursued by the Cognitive Science Dept. at UCSD and the Telecommunication and Computer Networks research group at the Univ. of Hamburg.

Section 2 provides more details on the inherent limitations of MOSs. TPMs are introduced in Section 3 as a prospective alternative. Some experimental results gained in the Coqos project are presented in Section 4, providing some early empirical evidence on relations between multimodal QoS parameters and cognitive human task performance. Section 5 concludes the paper.

*This research was supported in part by Deutsche Forschungsgemeinschaft (DFG) under award ME 1703/1-1.

2 A Critique of Mean Opinion Scores

MOS are usually obtained by *subjectively rating* stimuli with respect to a criterion like inter- or intra- media qualities in a presentation. Subjects express their judgements of media qualities according to a given scale. Finally, the scores are averaged across subjects to obtain the final MOS.

Due to its intraspective approach, MOS have certain inherent limitations that restrict its applicability. For example, Goodman and Nash [1] stated that:

> ... discrepancy [of MOS scores] may have to do with the impact of the distortion on Japanese speech or it may reflect differences in the way listeners use the rating scale.

We have identified some constraints where a MOS-based approach may fail to provide significant differentiation for adaptation or, even worse, may suggest misleading strategies. All cases could prove relevant in distributed multimedia environments. MOS tend to exhibit some or all of the following properties that could corrupt results:

1. Insensitivity to effects of unconsciousness;
2. Blurring of relevant details;
3. Ignorance of subjects' perspectives and intraspective positioning;

4. Allowance for ambiguous results.

Ad 1: Disturbances that are consciously not noticeable can by definition not be measured with MOS. If, for example, audio and video are out of lip-synchronization to some degree, people, while being completely unaware of, could perceive a different sound than actually uttered by a speaker if confronted which such a stimulus. It is a well-known cognitive fact that a wrongly perceived, or misheard, voice can be triggered by the lip motion seen. Such an effect is referred to as the McGurk effect [2]. A mean opinion score would clearly not be sensitive to it since the subjects have no means to verify their percepts. *Ad 2:* MOS is an indirect measure, thereby reflecting meta-cognition like 'How do I like this?'. So gradual differences that might be apparent to the subjects of the study might be lost due to individual differences in judging, mood, a priori estimates etc. There is no absolute scale that can be used for MOS. *Ad 3:* Subjects often approach a given problem from a certain perspective which is often unknown to the questioners. With respect to audio quality, for example, some subjects could be more concerned with understandability or others with tonal fidelity, depending on the prospective purpose.

Ad 4: Multimedia systems must include multimodal tuning. Considering the quote from [1], it might be hard to actually use the MOS score for future decision making for the different media since MOS do not provide any knowledge how the envisioned quality will affect the users.

3 Task Oriented Performance Measures

Task oriented performance measure take a different approach and expose the subjects to different levels of the stimuli (e.g., different frame rates) and objectively measure the outcomes. The performed task is related to a given context and the measured performance is thus relevant to an application that requires this task. Common tasks are, e.g., repetition, memorization of words or sentences.

This represents an operationalized direct way of dealing with the subjects' percepts such that the additional level of self-reflection is removed and validation of the obtained data is alleviated. By this approach, unconscious effects can be detected since they degrade performance. When small frame rates are responsible for McGurk effects, as described by Nakazono [5], we can measure the degradation in performance by wrong answers with the TPM-approach, whereas the MOS score might wrongly indicate a good presentation, a little jerky one perhaps. Instead of relying on users' opinon, TPMs provide an objective, yet individual, means to overcome limitations identified for MOS.

We envision standard task performance tests which can be universally applied to certain scenarios, achieving comparability and reproducability.

4 Some Empirical Results

4.1 Preliminaries

In our ongoing experiments, constraints on frame rates and on audio-visual skew as well as many interdependencies between these two factors are investigated. Some of intra-media parameters have already been proven elsewhere to affect human speech perception. However, only very few earlier studies have addressed the effect of how different frame rates and skews affect task performance. To our knowledge, no study has ever systematically addressed the interdependency of frame rate and audio-visual skew.

4.2 Outline of Experiments

The 15 subjects were mostly students of the University of California San Diego, between 18-32 years old. The experiment consisted of 8 blocks interleaved by breaks. Each block was made up of 60 stimuli, resulting in a total of 480 stimuli shown to each subject which took about 50 minutes. The task was to identify the second consonant in a four syllable nonsense words.

The words spanned all permutations of three consonants interleaved and headed by the vowel 'a' using the the four consonants 'b', 'd', 'g', 'v', e.g., 'adavaga'. This resulted in a total of 64 different stimuli words which were prepared with 30 different combinations of frame rates (30, 15, 10) and skews ($\pm160, \pm120, \pm80$ and 0ms) (a negative skew indicating that the audio is leading the video signal). Two of the consonants ('b','v') are labial, whereas ('d','g') do not require lip movement. The experiment follows a within-subject design, i.e. all subjects are exposed to all the different configurations of frame rate and audio-visual skew.

Sumby and Pollack [8] reported that the relative contribution of visual information is independent of the signal-to-noise ratio (S/N), but the absolute contribution could be more profitably exploited at low S/N. In order to explore the effects of frame rate, skew and the interactions between frame rate and skew, the audio signal was mixed with some amount of white noise, being about 11 dB louder than the signal. Therefore, the base performance was set to a level that ensured no clipping of the effects at the top of the scale.

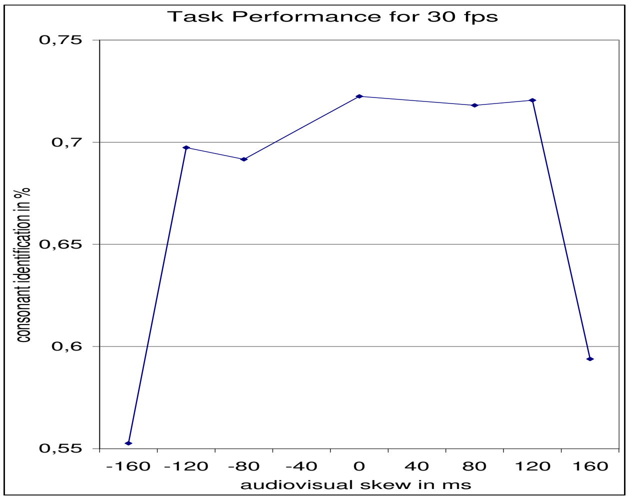
The experimental setup incorporated knowledge from earlier work, such as Massaro [3], Pandey et. al[6], McGrath et. al [4], and Nakazono [5].

4.3 Results & Discussion

Part of the results is depicted in the attached figure. A clear effect is evident for ±160ms where the performance drastically decreases. Whereas a positive skew seems to degrade the performance more gradually, a negative skew beyond 120 ms has a more abrupt effect.

In the study carried out by Steinmetz [7] the subjects had to detect an audio-visual skew and a MOS was defined as the level of annoyance realized by the subjects. The outcome can be referred to as based on a task-oriented measure. But it does not resemble tasks in the general and more *application oriented* sense, as we suggest. It is a well-accepted fact that speech perception is best for an audiovisual skew +80 ms (video leading audio). However, 30% of Steinmetz's subjects found that skew subjectively already annoying. Although more than 90%/70% of the subjects detected distortions in synchronization for +120/-120 ms, our experiments suggest that such de-synchronizations can often be tolerated as long as task performance is of concern.

The illustrated experiments indicate how TPMs come up with results that are more appropriate than those provided by MOSs. Explicit detection of skew is a very special task which may not be generalized to be applicable in other circumstances. The impact of distortions in synchronization can both be over- and underestimated with MOS, depending on the task at hand.



5 Summary

In the Coqos project Task Performance Measures and a corresponding framework are suggested and pursued as a novel and more suitable means for determining utility curves. TPMs are intended to avoid limits inherent in traditional measures like Mean Opinion Scores. MOS rely merely on subjective ratings rather than on more objective performance in relation to a particular task or application of interest. Informational relevance and its impact on subjects can be measured more effectively by TPMs. Inhibiting psychological and cognitive effects like consciousness or non-consciousness of degradations or individual focusing and perspectives of subjects can be more appropriately evaluated and dealt with by means of TPMs. The increasing importance of adaptation, in particular with the advance of MPEG4, as a means for QoS provisioning, both in wireless and wired environments, require sensible techniques to effectively determine utility curves.

References

[1] Goodman, D. J., Nash, R. D. – 'Subjective quality of the same speech transmission condition in seven different countries', IEEE Transactions on Communications, Vol. COM-30, No. 4, Apr. [1982]

[2] McGurk, H., MacDonald, J. – 'Hearing lips and seeing voices', Nature, vol. 264, (no.5588), 23 Dec., [1976]

[3] Massaro, D. W., Cohen, M. M., Smeele P. M. T. – 'Perception of asynchronous and conflicting visual and auditory speech', Journal of the Acoustical Society of America, 100 (3), Sep. [1996]

[4] McGrath, M., Summerfield, Q. – 'Intermodal timing relations and audio-visual speech recognition by normal hearing adults', Journal of the Acoustical Society of America 77 (2), Feb., [1985]

[5] Nakazono, K. – 'Frame rate as a QoS parameter and its influence on Speech Perception', Multimedia Systems, 6, [1998]

[6] Pandey, C.; Kunov, H.; Abel, S. M.– 'Disruptive effects of auditory signal delay on speech perception with lipreading', Journal of Auditory Research, Jan. 26 (1), [1986]

[7] Steinmetz, R. – 'Human perception of jitter and media synchronization', IEEE Journal on Selected Areas in Communications, vol.14, (no.1), IEEE, Jan., [1996]

[8] Sumby, W., and Pollack, I. – 'Visual contributions to speech intelligibility in noise.', Journal of the Acoustical Society of America, 26, [1954]

[9] Wolf, S., Dvorak, C. A., Kubichek, R. F., South, C. R., Schaphorst, R. A., Voran, S. D. – 'Future work relating objective and subjective telecommunications system performance', Proceedings IEEE Globecom [1991]