A NEW METHOD FOR IMMERSIVE AUDIOVISUAL SUBJECTIVE TESTING

Margaret H. Pinson[1]*, Marc Sullivan*[2]*, Andrew Catellier*[1]

[1]National Telecommunication and Information Administration; [2]AT&T

ABSTRACT

An immersive subjective test method is proposed in which subjects view each source stimulus only once. In order to encourage a subject's engagement with test content, longer stimuli are used. Distractor questions are used in addition to the traditional MOS scale in order to focus the subject on the intended application. A speech quality experiment is conducted with this method, and the results compared to those obtained with traditional methods. The consistent rank ordering among datasets demonstrates the validity of the immersive method.

1. INTRODUCTION

Current subjective quality testing standards, as Kortum and Sullivan [1] phrase it, "tend to reduce the human observer to the role of a simple detector." Subjects rate the same stimuli repeatedly—perhaps 25 times each—resulting in memorization, fatigue, and boredom. Fundamentally, it is unknown if this aspect of the experiment design biases the experiment results.

Sullivan and colleagues [1], [2] developed an alternative subjective testing method for measuring video quality. The idea is to more accurately assess video quality by immersing the subject in a naturalistic viewing experience. Each subject sees each stimulus only once. The pairing of scenes to HRCs[1] changes from one person to another. This focuses the subject on the system's intended usage scenario. The goal is to more accurately measure the system quality and acceptability.

This paper summarizes the immersive subjective testing method. To analyze this method, ITS performed an immersive subjective test of speech quality impairments. This test is described and the data collected compared to those obtained previously using traditional methods.

2. IMMERSIVE VIDEO QUALITY TESTING

2.1. Traditional Methods

There are a variety of subjective test standards with different goals. For example, ITU-T Rec P.910 and ITU-R Rec. BT.500 measure the quality of entertainment video, ITU-T Rec. P.800 measures speech quality, the modified rhyme test from ANSI S3.2 measures speech intelligibility, and ITU-T Rec. P.912 measures object recognition rates in video. These traditional methods ask multiple choice questions, as this simplifies both the task and the data analysis.

The modified rhyme test and ITU-T Rec. P.912 use artificial content (e.g., speech read from scripts, the same scene filmed multiple times with small differences). Content typically viewed or heard in a real-world situation contains redundant information and context clues. These context clues can be used to infer the correct answer. This makes it difficult to design multiple-choice questions with equally likely answers.

Though details differ, the basic structure of the traditional testing methods is quite similar. A small set of source stimuli are chosen for characteristics that exercise the encoder (e.g., phonemes for speech, spatial-temporal characteristics for video). Consequently, these stimuli are often artificial, like the Harvard balanced sentence "The hogs were fed chopped corn and garbage" or the ITU-R Rec BT.802 standard video sequence "Calendar and Mobile." Ideally, the stimulus set includes the full range of audio/video characteristics (e.g., the Harvard sentence sets include a balance of English phonemes).

Subjects are typically asked to rate every source and HRC combination. This maximizes measurement accuracy for each individual stimulus and allows a systematic comparison between all HRCs. These test methods isolate audio quality from video quality, unless the impairment to be tested involves both (e.g., audiovisual synchronization). Stimuli are short, often 6 to 10 seconds. The Single Stimulus Continuous Quality Evaluation (SSCQE) method from BT.500 presents the subject with a stimulus of long duration, but the subject is asked to move a slider on a rating quality scale every half second to rate current quality.

[1] Hypothetical Reference Circuit (HRC) is a fixed combination of a video encoder operating at a given bit-rate, network condition, and video decoder.

During a subjective test, the subject's task is to answer one question: what is the quality of this stimulus? Subjects are asked to ignore the content, yet we find that the content nevertheless influences people's perceptions of quality [3], [4].

2.2. Beyond Traditional Subjective Tests

We can observe a need for alternate subjective testing methods from the number of researchers who have modified or created entirely unique techniques. Four very different examples follow.

Staelens *et al.* [5] compared the impact of blockiness and frame freezes within full length movies with traditional subjective testing data. Some subjects watched the movie at home, then opened a sealed envelope and answered a questionnaire. Other subjects rated the impaired segments in the laboratory using ITU-R Rec. BT.500. Staelens found that the relative impact of impairment types changed with the setting. While watching the movie, the subjects were more tolerant of impairments that did not interrupt the flow of the movie.

Borowiak *et al.* [6] reversed the task. The subjects watched long videos (30 minutes), and the encoding quality occasionally dropped. When the subject noticed the drop in quality, they turned a knob to request a higher quality level. Turning the knob too far decreased the quality again. Subjects did not always return to the highest quality level despite the ability to do so. Instead, subjects seemed to choose an acceptable level of quality and return to that quality level.

The Staelens and Borowiak experiments indicate that the traditional subjective testing methods may not accurately predict the quality perceived by end-users. The experiment designs shift the focus from quality to the intended usage scenario through the use of long video sequences, audiovisual content, and novel stimuli. The drawback was that each subject participated for a long period of time yet relatively little data was gathered from each subject.

Cermak [7] conducted two surveys of digital cable subscribers. These surveys examined quality issues that traditional video quality experiments cannot, such as error stoppage (i.e., "An error message appears on the TV screen, and the video and audio stop. The cable box has to be reset." [7])

Krishnan and Sitaraman [8] analyzed data collected from a content delivery network. Using actual client data, Krishnan and Sitaraman estimated rates of abandonment, engagement, and repeat viewership in relationship to internet video impairments such as rebuffering.

The Cermak and Krishnan experiments demonstrate the need to ask different questions about video quality and acceptability in the context of a particular usage scenario. The drawback is that all control of the content and environment was lost.

2.3. Immersive Method

The immersive method includes several elements from these four papers. The intention is to maintain the ability to directly compare the quality of two different HRCs, yet put the subject in the frame of mind of using the system for its intended application:

- Enjoying a movie on TV
- Watching a YouTube video on a smartphone
- Talking with a friend on a video call

Longer stimuli are used to encourage this illusion and to engage the subject in the content matter (e.g., one minute). The content would ideally be interesting and consistent with content typical for that application.

The immersive method always matches the sensory experience of the target application—not the impairment modality. A video-only presentation poorly represents a user's experience of an audiovisual application. All immersive tests of broadcast video or video-on-demand applications present both audio and video, because consumers rarely watch videos with no sound. However, immersive tests of cell phones would present audio-only; and immersive tests of surveillance video would present video-only.

The use of audiovisual stimuli to evaluate video-only or audio-only impairments has consequences. The first is that subjects should always be asked to rate the overall audiovisual quality. Beerends and Caluwe [11] asked subjects to rate five aspects of the same stimuli:

1. The overall quality of the audiovisual stimuli
2. The audio quality of the audiovisual stimuli
3. The video quality of the audiovisual stimuli
4. The audio quality of the audio stimuli (audio only)
5. The video quality of the video stimuli (video only)

In this study, the subjects were not able to fully separate the audio quality from the video quality. Therefore, it is unreasonable to show subjects audiovisual stimuli and expect them to rate video quality only when using this immersive method. One can minimize the impact of audio quality on video quality by examining multiple video impairments while holding the audio quality constant (or vice versa).

The second impact of using audiovisual stimuli is that the range of mean opinion score (MOS) values will change. Pinson *et al.* [9] demonstrate that a multiplicative model (1) fairly accurately predicts audiovisual quality:

$$av \cong a \times v \qquad (1)$$

where a is audio MOS, v is video MOS, and av is audiovisual MOS. If only video quality is varied in an audiovisual test, it is reasonable to assume that a constant scaling factor and bias will identically impact all audiovisual MOS. The relative ranking of impairments should remain the same. We expect to see a scaling factor and bias when two different subjective tests are performed with traditional methods, because MOS is relative, not absolute [10].

The change from video-only (or audio-only) stimuli to audiovisual stimuli will impact our ability to distinguish between HRC MOS at some level of statistical significance. Pairing video-only impairments with constant quality audio will decrease the quality range and could cause saturation on the rating scale. Conversely, a greatly increased source stimulus pool will reduce the confidence intervals of HRC MOS. Whether the Student's *t*-test will be more or less sensitive is not known.

In the immersive method, each source stimulus is viewed or heard only once by each subject. The use of unique sources prevents the subject from memorizing the stimulus and avoids the boredom that often results from monotony. By showing test subjects each source only once, the influence of stimulus memorization cannot confound the results of the study.

Balance across the test is obtained by showing different combinations of sources and HRCs to each subject (or different sets of subjects, if multiple people view simultaneously). The number of sources should be an integer multiple of the number of HRCs. Preferably each subject should see five to ten stimuli for each HRC. This will yield a good estimate of each subject's opinion of each HRC (see Figure 3 of [3]).

Showing sources to subjects only once results in:

- A reduction in the quality measurement accuracy for each individual stimulus (e.g., "Calendar and Mobile" at MPEG-2 at 2 Mbps)
- An increase in the quality measurement accuracy for the HRC as a whole (e.g., MPEG-2 at 2 Mbps)

Given an immersive video test of *w* source stimuli, *y* HRCs and *n* subjects, the researcher will create every combination of source stimulus and HRC, for a total of (*wy*) stimuli. Each subject rates (*w*/*y*) of these stimuli for each HRC. When all subject scores are pooled, approximately (*n*/*y*) subjects will rate each individual stimulus, and all *n* subjects will rate each HRC.

The accuracy of the per-stimulus measurement (MOS) depends on the total number of subjects. This accuracy will decrease compared to traditional testing methods, because a subset of the subjects rate any given stimulus. For example, if *w*=30, *y*=5, and *n*=40, then around 8 subjects will rate each stimulus.

Where MOS is the average of all subjects for one stimulus, HRC MOS is the average of all source MOSs for a particular HRC. The HRC MOS accuracy depends primarily on the number of sources (*w*), not the number of subjects (*n*). The standard deviation of HRC MOS depends on how well the source stimuli represent the larger set of all available content. We reduce this standard deviation by increasing the number of source stimuli. Increasing the number of subjects has a minor impact when compared to the impact of increasing the number of source stimuli. Suppose we choose five source video stimuli depicting sports. Increasing the number of subjects will not improve our understanding of video generally, such as news, movies, adverts, cartoons, music video, sports and home video—it just increases our knowledge about those five sports videos.

The immersive method asks two targeted questions and three or four distractor questions. The first targeted question asks for the overall quality of the image and sound and is used to calculate MOS. The second asks for the subject's interest in the subject matter (i.e., their opinion of the content). This allows investigation of the influence of the source stimuli on the MOS data.

The same distractor questions are asked for all trials. Thus, the distractor questions must be generally applicable to all of the source stimuli. The distractor questions should be multiple-choice. These two constraints can aid in keeping the overall cognitive task simple.

The distractor questions serve two purposes. First, they focus the subject on the clip as a whole, instead of only the clip's quality [1]. Second, the distractor questions shift the subject's attention onto whether or not the stimuli would be acceptable for the particular application. The extent to which the distractor questions can measure acceptability will depend upon the experimenter's ability to pose appropriate questions. That topic is beyond the scope of this paper.

Unlike traditional methods, the immersive method has a minimum possible number of subjects: one for each HRC to be examined. AT&T has observed stable results when using 30 to 40 subjects to rate four or five video-based HRCs.

A potential drawback of the immersive method is the small number of HRCs in each test. The immersive method uses long video stimuli, five or six questions, and four to six stimuli for each HRC. Thus, the total number of HRCs in a given test should be reduced, to prevent impossibly long tests. A possible solution might be to greatly increase the number of subjects and have each subject rate only one or two stimuli for each HRC. Such a large increase in the number of subjects would be prohibitively expensive unless crowdsourcing is used.

3. IMMERSIVE SPEECH QUALITY TEST DESIGN

To evaluate the immersive subjective test method, ITS designed and conducted an experiment on speech quality. This topic was chosen due to the availability of:

- A large set audiovisual footage that contains a wide variety of people speaking
- Prior publication of subjective speech quality ratings
- POLQA's objective speech quality ratings to serve as a third comparison (Perceptual Objective Listening

Quality Assessment is a full reference objective voice quality algorithm described in ITU-T Rec. P.863.)

Speech quality subjective tests tend to have a more narrow range of quality than video quality subjective tests [9]. This relatively narrow range of quality serves as a challenge for the immersive subjective test concept.

The source stimuli depict a variety of people of different ages, genders and ethnicity. They are discussing various topics in response to an interviewer, off screen. The audio track contains a single person talking in English, using natural (though occasionally stilted) speech patterns. The source audio is usually pristine, however half of the stimuli have soft background noise (e.g., from an air conditioner) and one source has a small amount of clipping. The audio was converted to mono and normalized to -26 dB below clipping. The beginning and ending of each audio was ramped from or to silence.

Each video depicts one person's head and shoulders in TV interview format, with a mottled gray background. The video was filmed in 1080i 59.94 fps on a variety of broadcast quality camcorders, de-interlaced and converted to 1080p 29.97 fps for presentation on a laptop.

Twenty source stimuli were selected for the test, and two for the training session. The source set includes two stimuli from each of five males and five females. Two stimuli from a sixth female are used for the training session. The stimuli range from 34 to 52 seconds long, with an average length of 42 seconds. Each stimulus conveys a segment of speech that can be understood in isolation (i.e., without the context of the prior interview footage). The different durations reflect the length of time required to present a complete thought, though maintaining a constant video length would be preferred (and perhaps optimal).

Four audio impairment levels were selected:

- A1 is AMR narrowband, mode 0 (4.75 kb/s)
- A2 is AMR narrowband, mode 7 (12.2 kb/s)
- A3 is AMR wideband, mode 1 (8.85 kb/s)
- A4 is AMR wideband, mode 8 (24.0 kb/s)

The impairment levels A1-A4 were chosen because we had access to MOS values from prior subjective tests (see section 4.1). The coded audio stimuli were all normalized to -26 dB below clipping and time shifted to ensure audiovisual synchronization within ±1 ms (according to POLQA's comparison between the original and coded speech). One training stimulus was compressed to level A1 and the other was compressed to level A4.

The following five multiple-choice questions were posed to the subjects. Each question is followed by the allowed answers.

1. What topic was this person discussing?
   Occupation, traveling, family, self, memories, other
1. How interesting did you find this clip?
   Intriguing, interesting, neutral, uninteresting, boring
2. What attracted your attention the most?
   Message, clothing, face, gestures, manner of speaking
3. Would you enjoy having a conversation with this person?
   Very likely, somewhat likely, neutral, unlikely, very unlikely
4. How would you rate the overall quality of the sound and picture?
   Excellent, good, fair, poor, bad

The subjective test was performed on a 17" laptop using an updated version of the automated software used by Catellier *et al.* [12]. The video was lightly compressed using H.264/AVC to ensure reliable playback, and the audio compression was transparent (see [12] for details). The test was conducted in one session.

Data were then gathered from 16 subjects who were recruited through a temporary employment agency. The agency was asked to supply people with good vision and hearing. Subjects were not screened for vision or hearing. Each of the 16 subjects rated a different combination of sources and HRCs. The combinations were chosen such that similar numbers of subjects would rate each impaired stimulus (i.e., source stimulus × HRC combination), and each subject would observe and rate a different subset of the source stimuli for each HRC.

The experiment sessions were conducted in a sound-isolated room with background noise measured below 20 dBA SPL. Philosophically, the immersive method is better suited to a simulated living room environment (such as proposed in [12]). However, using the sound-isolated room allows for a more direct comparison to previous tests using the same audio impairment levels. The sound was delivered using circumaural headphones (with a specified -3 dB bandwidth from 16 Hz to 30 kHz) and the laptop's internal sound system. The room lighting used full spectrum light bulbs and dim light levels.

4. DATA ANALYSIS

4.1. Comparison to Traditional Testing Methods

The immersive speech quality data will be compared with three different sets of speech quality measurements.

The first dataset predicts the quality of our speech samples with ITU-T Rec. P.863 (POLQA). POLQA is an objective speech quality model approved by the ITU-T in 2011. POLQA is the successor of ITU-T rec. P.862 (PESQ), which was approved in 2001. Note that ITU-T Rec. P.863 recommends the use of POLQA only for stimuli that contain no more than six seconds of active speech.

The second dataset is a subjective test performed by Voran and Catellier [13], partially to investigate speech quality delivered by modern speech codecs. This experiment was conducted according to ITU-T Rec. P.800 using the absolute category rating (ACR) scale with 5 levels. This same ACR scale is presented in question 5 of the immersive speech experiment described in the previous section. The experiment included 36 speech samples in English from two females and two males. The audio was recorded in a sound-isolation chamber with studio-quality recording equipment, and thus had no background noise. The speech ranged in duration from one to five seconds and simulated typical telephone conversation talk-spurts. The Voran experiment and the immersive speech quality experiment were both performed in the same sound-isolation chamber, using the same headphones.

The third dataset is a subjective test performed by Ramo [14] to compare a wide variety of different audio codecs. This experiment was conducted according to ITU-T Rec. P.800 using ACR modified to have 9 levels. Level 9 was labeled "excellent", level 1 was "very bad", and 2-8 had no labels. For comparison purposes, these scores were mapped from [1..9] to [1..5] using the mapping:

$$\hat{s} = \frac{s}{2} + \frac{1}{2} \qquad (2)$$

4.2. Analysis of Results

The total duration of all the immersive stimuli was 14.06 min. The subjects took from 17 to 40 min to complete the test session of 20 clips, with an average of 24 minutes.

The R^2 statistic indicates that subject matter interest (question 2) explains 10% of the spread of subjective scores (question 5). However, that influence impacts all HRCs equally. None of the other distractor questions were evaluated.

The immersive test's HRC MOS is calculated as an average of question 5 for all subjects and stimuli. The HRC MOS values for all four datasets are listed in Table 1 and displayed as a bar graph in Figure 1 top. The datasets are mapped to POLQA and displayed in Figure 1 bottom. The linear fits are as follows:

$$y_{immerse} = -7.20\,x - 2.87 \qquad (3)$$
$$y_{voran} = 0.63\,x - 0.79 \qquad (4)$$
$$y_{ramo} = 0.89\,x - 0.89 \qquad (5)$$

Table 1. HRC MOS for A1-A4

	A1	A2	A3	A4
Immersive	3.42	3.70	3.76	3.89
POLQA	2.66	3.33	3.56	4.04
Voran [13]	2.54	3.42	3.83	4.22
Ramo [14]	2.03	2.70	2.95	3.58

Table 2. Pearson Correlation between Datasets

	Immersive	POLQA	Voran	Ramo
Immersive	1.00	0.99	0.99	0.98
POLQA	0.99	1.00	0.99	0.99
Voran [13]	0.99	0.99	1.00	0.98
Ramo [14]	0.98	0.99	0.98	1.00

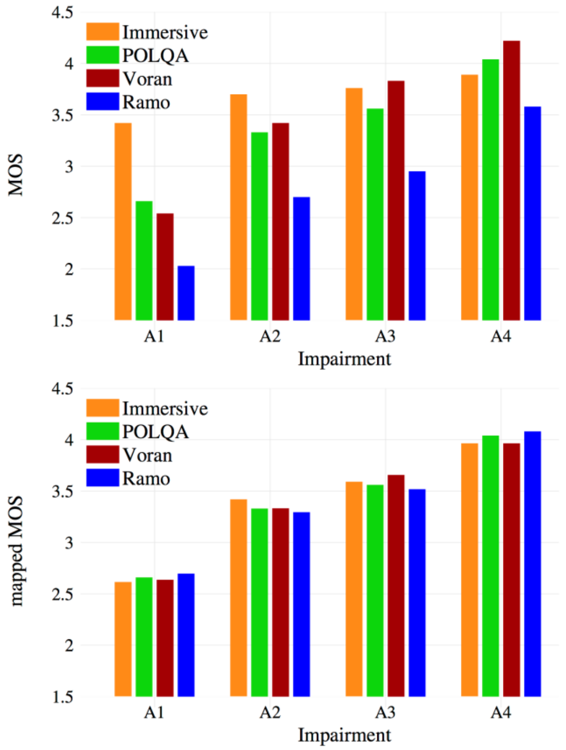

Figure 1. Bar graph compares HRC MOS for all four datasets (top) and dataset mapped to POLQA (bottom).

The data from the Immersive test, Voran, Ramo and POLQA agree with each other closely. The HRC MOS values for Ramo [14] are half a unit lower than POLQA and Voran, possibly due to the use of a 9-level scale. An adjustment of equation (2) would remove this offset. Table 2 shows that the Pearson correlations between each pair of datasets are all similar (0.98 to 0.99).

The HRC MOS from the immersive test agree with the prior datasets, in that the ordering and relative MOS distributions match. The influence of the high quality video can be seen in the narrow range of immersive MOS scores, and the shift of those scores toward the upper end of the scale. The Student's *t*-test at 95% confidence indicates (A1,A3) and (A1,A4) were statistically different for the immersive data, while all HRCs were statistically different for the Voran data. The confidence intervals reported by Ramo indicate all HRCs are statistically different.

5. CONCLUSION

The immersive subjective method was shown to replicate results of prior speech quality experiments conducted with traditional methods. The immersive HRC MOS values differed by a gain and offset, which can be explained by the presence of high quality video. The immersive method cannot replace traditional methods yet has promise for some problems.

This immersive method has potential for applications that are difficult to analyze with traditional subjective testing methods. By drawing on techniques used to design questionnaires, the distractor questions could be used to infer the minimum level of quality that is acceptable for a particular application. The obvious application is commercial decisions on video products and services, where the vendor needs to decide between perceived quality and cost. A second application is video systems for sign language, where the layered interaction between different linguistic elements makes it difficult to create artificial stimuli for an ITU-T Rec. P.912 style task oriented experiment. A third application is audiovisual communication for emergency telemedicine applications. Immersive testing could help address the tradeoff between wireless bandwidth limitations and audiovisual quality in a situation where immediate action must be taken.

Crowdsourcing tests might benefit from the using the immersive method instead of traditional methods. Keimel *et al.* [15] observed that crowd-based subjects cannot be depended upon to complete an entire subjective test. The logical crowdsourcing task would be to rate one SRC for each HRC. Design balance could be maintained even if only one task is performed by the subject (e.g., each subject is given a different subset of scenes from a large scene pool). The distractor questions might provide an alternate mechanism for analyzing subject reliability, which is another problem identified by Keimel.

The immersive method has advantages for the subject. Even a 20 minute session using traditional methods is tiring, and subjects occasionally express dread at the prospect of the 2nd or 3rd such session. Expert subjects at ITS were more comfortable after a 20 minute immersive test session than after prior 20 minute sessions conducted using traditional subjective testing methods. They felt able to continue immediately. Researchers at AT&T have observed people leaving immersive sessions in good humor (e.g., a group of subjects laughing as they left a test focused on football content).

The audiovisual HDTV project of the Video Quality Experts Group (VQEG, www.vqeg.org) is interested in evaluating the immersive subjective testing method. Presentations of subjective tests performed with this method would be welcome.

7. REFERENCES

[1] P. Kortum and M. Sullivan, “Content is king: the effect of context on the perception of video quality,” *Proceedings of the Human Factors and Ergonomics Society 48th annual meeting*, 2004.

[2] M. Sullivan, J. Pratt and P. Kortum, “Practical issues in subjective video quality evaluation: human factors vs psychophysical image quality evaluation,” *First International Conference on Designing Interactive User Experiences for TV and Video (UXTV ‘08)*, Oct. 22-24, 2008.

[3] M. H. Pinson, M. Barkowsky and Patrick Le Callet, “Selecting scenes for 2D and 3D subjective video quality tests,” *EURASIP Journal on Image and Video Processing*, Aug. 2013.

[4] P. Kortum and M. Sullivan, “The effect of content desirability on subjective video quality ratings,” *Human Factors*, vol. 52, no. 1, 2010.

[5] N. Staelens *et al.*, “Assessing quality of experience of IPTV and video on demand services in real-life environments,” *IEEE Transactions on Broadcasting*, Vol. 56, Issue 4, Dec. 2010.

[6] A. Borowiak, U. Reiter and U. P. Svensson, “Quality evaluation of long duration audiovisual content,” *Consumer Communications and Networking Conference (CCNC)*, Jan. 2012.

[7] G. W. Cermak, “Consumer opinions about frequency of artifacts in digital video,” *IEEE Journal of Selected Topics in Signal Processing*, vol. 3 no. 2, Apr. 2009.

[8] S. S. Krishnan and R. K. Sitaraman, “Video stream quality impacts viewer behavior: inferring causality using quasi-experimental designs,” *ACM Internet Measurement Conference*, 2012.

[9] M. H. Pinson, W. Ingram, and A. Webster, “Audiovisual Quality Components: An Analysis,” *IEEE Signal Processing Magazine*, vol. 28, no. 6, pp. 60-67, Nov. 2011.

[10] M H. Pinson *et al.*, “The Influence of Subjects and Environment on Audiovisual Subjective Tests: An International Study,” *IEEE Journal of Selected Topics in Signal Processing*, Vol. 6, No. 6, Oct. 2012.

[11] J. G. Beerends and F. E. de Caluwe, “The influence of video quality on perceived audio quality and vice versa,” *J. Audio Eng. Soc.*, vol. 47, no. 5, 1999.

[12] A. Catellier, M. Pinson, W. Ingram and A. Webster, “Impact of Mobile Devices and Usage Location on Perceived Multimedia Quality,” *International Workshop on Quality of Multimedia Experience (QoMEX)*, Jul. 2012.

[13] S D. Voran and A. A. Catellier, “When should a speech coding quality increase be allowed within a talk-spurt?” *2010 IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, 2013.

[14] A. Ramo, “Voice quality evaluation of various codecs,” *2010 IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, 14-19 Mar. 2010.

[15] C. Keimel, J. Habigt and K. Diepold, “Challenges in crowd-based video quality assessment,” *International Workshop on Quality of Multimedia Experience (QoMEX)*, Jul. 2012.