Bachelorthesis Psychology

F. van der Sluis

Synergy effects with mobile (audio and) video telephony

Dr. R. van Eijk
Supervisor on behalf of:

Telematica Instituut

Dr. E.L. van den Broek
Supervisor on behalf of:

Universiteit Twente
de ondernemende universiteit

Telematica Instituut

Universiteit Twente
de ondernemende universiteit

Synergy effects with mobile (audio and) video telephony

F. van der Sluis (s0029491)
f.vandersluis@student.utwente.nl

ABSTRACT
Despite the progress in technology, mobile video telephony (MVT) is not used on a large scale. In a quest to explain the latter, we adopted a user-centered instead of technological perspective, where the influence of screen size on the synergy of audio and video was under investigation. 54 participants conducted an experiment in which the intelligibility of a standardized video-listening test was determined for three screen sizes: mobile phone, PDA and PC monitor. A significant increase in intelligibility for the large compared to the small screens was found. Moreover, a signal-to-(white) noise ratio of -9dB significantly limited the intelligibility of the videos. With respect to the Quality of Service of MVT, two conclusions result: 1) the display size should be maximized and 2) already limited amounts of noise decreases intelligibility. Consequently, we emphasize the need for both technological development and user-centered research on MVT.

1. INTRODUCTION

Person-to-person mobile video telephony is nowadays available to many consumers. The growing availability of 3G networks and services and the emerging features of mobile video phones enable a greater connectedness between its users, enhancing the social and emotional aspects of communication. Mobile video telephony is used for functional talk (22%), showing objects (28%), and social and emotional small talk (50%) [13].

Compared with traditional telephony, video telephony is enriched with video and, thus, facilitates multimodal perception. In last decades, the benefits of multimodal interaction became more and more apparent. Consequently, the interest in multimodal perception increased significantly, from both industry and science; e.g., visual-auditory modalities are complementary on individual phonetic features [15] and are synergetic: combined auditory-visual perception is superior to perception through either audio or vision alone [6]. This synergy is especially salient in noisy surroundings, where the bimodal advantage can become as large as a 39% increase on intelligibility [14].

Multimodal communication also influences memory and emotion; e.g., adding visual gestures to auditory speech improves the quality of the memory for speech [10]. Moreover, both modalities supply complementary information about emotions, which are effectively combined [3]. The latter explains the main use of mobile-video telephony for social and emotional talk. More generally, it illustrates the superiority of multimodal communication over unimodal communication.

Seen within the context of the mobile device, some restrictions may be imposed on these bimodal advantages. Several factors are expected to be of influence, among which are 1) bimodal issues: a) spatial coherence (or ventriloquism effect) [19], b) source coherence; i.e., two sources behave in a way that a heard stimuli is ascribed to them both [12], c) temporal coherence; i.e., different events take place at (almost) the same time and are thus seen as one stimuli [e.g., 5], and

d) visual dominance [4], 2) auditory issues [13], and 3) visual issues (i.e., temporal frequency, spatial resolution, screen size, and noise and zoom-level). Most of these topics have been researched. For example, temporal frequency and spatial resolution can be at low levels before decreasing intelligibility: a 25Hz temporal frequency [18] and 32x32 spatial resolution [1]. Except for screen size, most of these effects have been topic of extensive research.

Surprisingly, research on the effect of screen size on human multimodal perception is absent in mobile video telephony, this despite their small screen is one of its salient features. Hence, the major limitations in usage of video telephony are not merely technology related. The usage of the telephones' small screen is possibly a factor of importance. As attention focused on large screens, such as used for virtual reality and/or entertainment applications; small screens received little notice. In general, larger screens influence values such as arousal, sense of presence, attention and memory, and connectedness [7]. For most of these values the effects can be summarized as intensifying the values. Hence, "the larger, the better" seems to hold.

Screen size is best described by Field Of View (FOV), Pixel Per Inch (PPI), resolution, and the actual physical display size. FOV is the size of a screen relative to the eye of the user, taking the distance from the eye to the screen and width of the screen as input. This is related to the display size, as display size stands for the width and height of a screen. Resolution has been found to be of less importance for low-level effects such as intelligibility. A resolution of as little as 32x32 pixels was enough to enhance the bimodal intelligibility [1]. Resolution is related to PPI, as PPI can be computed from the display size and its resolution. Finally, concerning the physical size of screens, larger screens were found to improve performance [9].

This research examines the influence of screen size on the synergy levels of bimodal communication. Expected is that a decreasing screen size reduces the intelligibility of a message presented auditory as well as visually.

To answer this hypothesis, the importance of the visual modality relative to the auditory modality is increased by adding noise to the auditory channel. Consequently, changes in the visual channel have a greater effect on the final intelligibility.

2. METHOD

To study the influence of screen size, a within-subjects design has been used evaluating the effects of screen size, videos and their sequence. This gave a total of 36 different conditions (3! x 3!), based on the possible number of combinations of the three screen sizes and the three videos.

2.1 Participants

A total number of 54 participants voluntarily participated in the research. Since the required number of participants was a plural of 36 this was insufficient to fully counterbalance the possible confounding effect of sequence of presentation, yet this was found to be of no influence as shown further on at the Results section.

The age of the participants ranged from 18 to 28 years with an average of 20.3. 63% of the participants had the Dutch nationally, 37% had the German nationality. 96.2% of the participants judged their level of English as either good or reasonable. All participants had a (corrected to) normal vision and hearing.

2.2 Material: Listening test

To evaluate the participants' knowledge of the English language, an English listening test preceded the video-listening test. The listening test was part of the English listening exams on University preparatory education in the Netherlands.

The listening test was constructed by CITO (Central Institute for Testing). The test consisted of twelve parts of an interview with a probation officer about his job. After each part a multiple choice question, as were provided by CITO, had to be answered to test the participant's comprehension. These questions were standardized and scored with norm based correction forms.

2.3 Material: Video listening test

A set of nine videos with accompanying questions were selected from the original material provided by CITO. After each video an English multiple (three) choice question had to be answered to evaluate the intelligibility of the shown video. No restriction was made on the answering time. The test results were rated with the original CITO scoring forms, as measure for its intelligibility.

The video-listening test consisted of an interview with an exchange student. Most of the time, the face of either the student or the interviewer were shown when they were talking; though, sometimes video parts were shown of life in Africa. The videos were selected such that the face of the person talking was visible most of the time.

In order to increase the internal validity of the findings, three questions were left out of the analysis due to several invalidating factors. The videos of the left out questions showed irrelevant or false footage at critical moments and gave conflicting results concerning accuracy and answering time scores. Critical moments are those moments in which other footage is shown while the information needed to answer the question is told, resulting in a loss of synergy effects. Furthermore, false footage contradicting the answer outside critical moments was found to invalidate the results as well, an effect also known from other research [8].

Figure 1. Experimental setting.

2.4 Apparatus

A computer with a 15" screen and a headset was used to present the videos with audio or the audio alone.

The experiment took place in two rooms of a behavioural sciences research laboratory at the University of Twente. In each of the two rooms, the experimental placing was exactly the same. This experimental setting is displayed in Figure 1.

One of the most notable features of the experimental setting is the rope construction surrounding the participant's head, forcing the participants to keep a particular distance and position to the screen. In situations where participants are able to choose their own distance and position to a screen, they are likely to move closer when a small screen size is shown. Being closer to the screen makes the percept of the screen relatively larger and, thus, increases the FOV, which might diminish any effects of difference in intelligibility between the three screen sizes. The used ropes formed a square on forehead height. In addition, the chair and keyboard were also placed at a fixed position, which relieved the effort for the participants to fixate their head within the rope square. Figure 1 illustrates this setting; please note the ropes that make a square around the participant's head.

Three screen types were used, namely; computer screen size, PDA screen size and mobile (video) phone screen size, as specified in Table 1 together with the size of the screens (expressed as the length of the diagonal in centimetres), resolution of each screen size, and the FOV. In literature, different definitions of FOV can be found. In

Table 1. Screen type, size, resolution and Field of View (FOV) of each screen used in the experiment.

Screen type	Size (cm)	Resolution (pixels)	Field of View (°)
Computer	38.10	1280x960	13.48
PDA	12.82	394x316	4.50
Mobile (video)	6.02	177x158	1.12

this research, the FOV is defined as the angle subtended from the eyes to the left and right edges of the displayed screen [17].

All these screen sizes were shown on the same screen with a resolution of 1280x960. To prevent any confounding effects of an increasing video quality, the amount of pixels of the videos were kept equal but spread over a larger part of the display. Thus, reducing the spatial data density but enlarging the FOV.

2.5 Determination signal-to-noise ratio

For all videos, a SNR of -9dB was used to maximize synergy effects. This SNR is based on findings from several studies, indicating an increase in intelligibility from -6dB to -30dB [e.g., 6, 16]. Compared to these studies, the SNR has been kept low, because the used stimuli are more complex and longer than those used in the mentioned studies. Furthermore, a pilot study has been performed (N=6) to test the synergy of three SNR values (-6dB, -9dB and -12dB), showing all three SNR values had a reasonable and comparable effect size. The SNR was computed as follows:

$$\text{(1) } SNRdB = (RMSamplitude,signal - RMSamplitude,noise),$$

where the Root Mean Square (RMS) amplitudes of the signal and noise are, respectively, -15dB and -6dB and defined by

$$\text{(2) } RMSamplitude = 20*\log 10\ \frac{X}{Xref},$$

where X is either the power of the signal or the noise and $Xref$ is the power of the reference point of the used Decibel scale. For all mentioned dB values, the used scale is dBFS (Decibel Full Scale). For this scale the reference point is the maximum output level of the hardware. Next, the SNR was defined as a similar logarithmic function of the power of the signal divided by the power of the noise, and has been translated to the first equation.

Table 2. Accuracy scores on each question for each screen size, including norm scores from optimal conditions.

	Accuracy				
	Screen Size				
Question	Small	Medium	Large	Mean	Norm
1.1	0.44	0.44	0.56	0.48	0.91
1.2	0.44	0.56	0.72	0.57	0.67
2.1	0.28	0.39	0.33	0.33	0.98
2.2	0.72	0.78	1.00	0.83	0.98
3.1	0.61	0.83	0.83	0.76	0.87
3.2	0.94	0.83	1.00	0.93	0.66
Mean	0.57	0.64	0.74	0.65	0.85

Note. Norm scores adopted from CITO.

2.6 Procedure

Before the experiment, participants were told they were going to undertake a listening and a video-listening test for which they should remember as much as possible from the video. Furthermore, they were told to keep their head within the rope construction and that the experimenters would check on this using an installed video camera.

The experiment started with some questions concerning general demographic data; i.e., name, sex, age, occupation, and nationality. The second part contained the English listening test, testing the English level of the participants. The third part contained the three videos in three different screen sizes, as defined by one of the 36 possible conditions. Finally, the fourth part asked some questions about the experience of the participants with the experiment. The total duration of the experiment was approximately 20 minutes.

3. RESULTS

The descriptive statistics of the questions are shown in Table . A one-tailed t-test showed a significant difference between the average norm results per question of the CITO (M=0.85, SD=0.11) and the current results for the large screen size (M=0.74, SD=0.16),

t(53)=-2.83 (p<.01), indicating a significant influence of the used SNR.

All further described analyses are on video level (each video consists of two questions). The descriptive statistics are shown in Table , and pictured in Figure 2. The primary effect of screen size on intelligibility was analyzed using a Multivariate Analysis of Variance (MANOVA) on accuracy and answering time by screen size, video, and sequence for all videos. The MANOVA of accuracy by screen size showed that a small screen enhances intelligibility less than a large screen (F(2, 135)=4.60, p<0.05). However, one-tailed t-tests revealed no significant results on the comparison between small and medium (M=0.13, SD=0.12), where the difference between medium and large (M=0.20, SD=0.12; t(53)=2.60, p<.05), and between small and large were significant (M=0.33, SD=0.12; t(53)=4.27, p<.001). The correlation between screen size and accuracy was .21 (t(52)=3.00, p<.01). For answering time, no effects were found by screen size, which clearly shows from the data as well (see Table).

The different videos differed significantly in difficulty (F(2,135)=18.36, p<.001) and in answering time (F(2,135)=22.27, p<.001), as revealed by a second MANOVA. The influence of sequence was non-significant, indicating that there was no learning effect within the different trials each subject performed. Post-hoc Bonferroni tests showed

Table 3. Accuracy scores and answering times on each video for screen size, video and sequence.

	Mean (Standard Deviation)	
Factor	Accuracy	Answering Time (s)
Screen Size (FOV)		
Small (1.12°)	1.15 (0.66)	30.33 (11.73)
Medium (4.50°)	1.28 (0.66)	29.53 (14.02)
Large (13.48°)	1.48 (0.57)	30.46 (13.21)
Video		
1	1.06 (0.68)	32.76 (12.23)
2	1.17 (0.57)	36.05 (11.60)
3	1.69 (0.47)	21.51 (10.32)
Sequence		
First	1.31 (0.70)	30.14 (12.87)
Second	1.28 (0.63)	30.80 (13.64)
Third	1.31 (0.61)	29.37 (12.51)
Mean	1.30 (0.64)	30.11 (12.95)

significant differences (p<.001) between video 1 and video 3 (M=0.63, SD=0.11) as well as video 2 and video 3 (M=0.52, SD=0.11); hence, video 3 was experienced as more difficulty than the other two. The primary effects of screen size, video and sequence can be viewed graphically in Figure 2.

The English Level was found to be sufficient for all subjects (M=8.89, SD=1.98). Furthermore, it did not correlate with accuracy scores on the video test, r(54)=.04 (p>.05), indicating that English Level did not influence the performance on the tests. Finally, gender did not influence the correlation between screen size and accuracy. For both men and women the correlation remained .21, though being a trend for men (t(20)=1.81, p<.07) and significant for women (t(30)=2.29, p<.05).

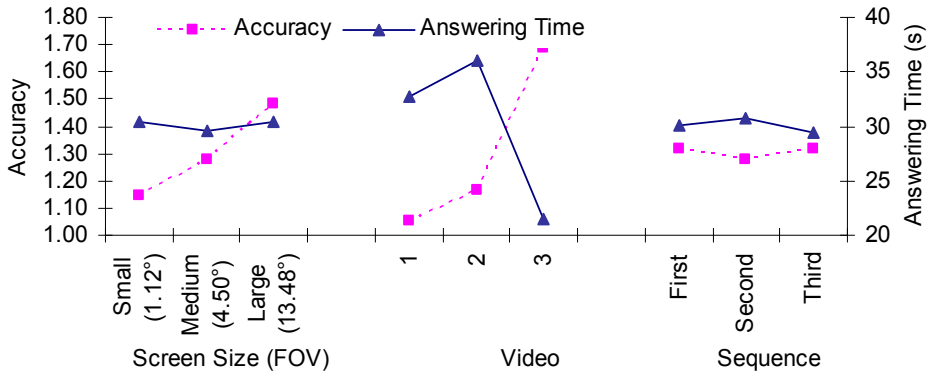

Figure 2. Accuracy and answering time results shown for each screen size, field of view (FOV), video and sequence.

4. DISCUSSION

Already more than half a century ago, Sumby and Pollack [16] described the synergy of our auditory and visual percepts; i.e., seeing someone speak helps hearing him. Par excellence, it illustrates the holistic process underlying human multisensory perception and with that, the potential of mobile video telephony. Hereby, the synergy between audio and video is especially of interest. This since mobile video telephony suffers from a variety of sources of noise and leaps in the signal and, as Sumby and Pollack [16, p. 214] stated: "the visual contribution to speech intelligibility [...] increases as the speech-to-noise ratio is decreased". Regrettably, studies such as that of Sumby and Pollack [16] are rare, in particular within the context of mobile video telephony.

In line with the work of Sumby and Pollack [16], the main hypothesis of this study stated that the intelligibility of a message presented visually as well as auditory reduces when the screen size is reduced. This was confirmed by a significant difference in accuracy scores on the video-listening test for each different screen size.

In a setting where through a SNR of -9dB any effects of the quality of the visual stimuli are enhanced, the effect of screen size showed with a relatively small amount of participants. The correlation between screen size and accuracy scores is .21, indicating screen size is indeed an influential factor in intelligibility.

This effect showed only for accuracy scores on the video-listening test, not for answering time measurements. This was contrary to the difference in difficulty of the different videos, which showed on both measurements. Where it was expected answering time is somehow related to the difficulty of processing the presented stimuli, this indicates a different level of processing for the content of the video and the integration of the different modalities. Furthermore, these results showed the effect was not influenced by any learning in the short run (the tests took about 12 minutes). Also, the effect was found to be unrelated to differences in comprehension of the English language, given that there was a very homogenous population, all at least reasonable capable of understanding English.

Several factors might restrict the applicability of these findings, concerning the used stimuli and tests. The external validity of the results is somewhat reduced by the use of a non-natural SNR. The effects might be different at a natural SNR, and might be more salient at other cognitive levels (e.g., emotional). In real conversations using mobile video telephony probably other characteristics of the message might be important than are tested with the used stimuli. Therefore in future research more natural stimuli might be used; e.g., simulating a real phone conversation instead of the in the described experiment used listening task.

Across different studies, gender differences are consistently found: women tend to react more to differences in screen size than men [7]. However, results of this study indicate none or very minor differences between men and women for the described effects. The only difference was in effect of screen size on accuracy between men and women, having a trend for men and significant effects for women. Since the correlation did not change between the groups, this difference in effect can even be attributed to difference in group sizes.

As shown is intelligibility higher with a bigger screen size when a noisy auditory stimulus is available. In practical use, mobile video telephony takes place with a small screen and probably with a less professional headphone or even without a head phone. This can negatively influence the intelligibility even more. A higher quality sound output for mobile (video) phones could limit the loss of intelligibility occurring with small screen sizes, although this still restricts the larger benefits of bimodal communication. In order to benefit fully from bimodal communication, larger screens should be used.

Recent technological developments relieve the problem of using large screens with mobile video telephony. Two of the more interesting

products are lightweight high-resolution video glasses [2] and flexible electronic paper, which is now available in A4 size in color [11]. Both applications have their own advantages and disadvantages. The video glasses disconnect the user from its surroundings, which can cause problems in communication. The electronic paper has a limited resolution and supports a maximum of 4,096 colors. On the other hand, the electronic paper can be viewed from a full 180 degrees, so that images always appear crisp, even when the display is bent. Alternatively, infrastructural solutions can also be sought to relieve the problem. For example, Bluetooth enables the connection of mobile devices with large screens that are available at that moment. Then, video telephony can be performed using a large display.

Our findings show some new and unexpected directions for future research on improvement in Quality of Service of mobile video telephony. It shows that the effects of audio and video quality on perception and cognition cannot be treated separately and that any significant improvement can only be derived when their synergetic effects are taken into account. Furthermore, our study indicates that there is a threshold for which video does not contribute to the intelligibility of the audio. More research need to be done to find out the conditions under which this threshold is reached and, subsequently, the actions to be taken on the mobile device or in the network to deal with this event.

This study is rare in its kind since it stressed usability research in a field dominated by technology. It places fundamental work on human perception and information processing in the context of the field of mobile video telephony. The experiment conducted, revealed one of the possible reasons for a large scale success of mobile video telephony: the limited synergy of audio and video with small screens. In parallel, this multimodal feature of mobile video telephony revealed to be very useful in case of noisy signals, as occur with mobile video telephony. With that, both the vulnerability and strength of mobile video telephony are illustrated, which emphasize its fragile future.

5. ACKNOWLEDGEMENTS

The CITO, with in particular Jan van Thiel, is gratefully acknowledged for their generous cooperation in selecting and, subsequently, preparing suitable video-listening tests. In addition, we thank Ronald van Eijk, Johan de Heer and Sorin Iacob of the Telematics Institute for their cooperation and fruitful discussions during this study. Last, we thank all subjects for their voluntary participation in this study.

6. REFERENCES

[1] Brooke, N. M. and Templeton, P. D. Visual speech intelligibility of digitally processed facial images. Proceedings of the Institute of Acoustics, 12, 1990, 483-490.

[2] BV RelaxView. URL: http://relaxview.nl/ [Last accessed on June 09, 2007].

[3] De Gelder, B. and Vroomen, J. The perception of emotions by ear and by eye. Cognition and Emotion, 14, 2000, 289-311.

[4] Dijkstra, A. A computer model for bimodal sublexical processing. Swets & Zeitlinger, 1994.

[5] Dixon, N. F. and Spitz, L. The detection of audio-visual desynchrony. Perception, 9, 1980, 719-721.

[6] Erber, N. P. Auditory-visual perception of speech. Journal of Speech and Hearing Disorders, 40, 1975, 481-492.

[7] Grabe, M. E., Lombard, M., Reich, R. D., Bracken, C. C. and Ditton, T. B. The role of screen size in viewer experiences of media content. Visual Communication Quarterly, 6, 1999, 4-9.

[8] Gruba, P. The role of video media in listening assessment. System, 25, 3, 1997, 335-345.

[9] Josephs, R., Giesler, R. and Silvera, D. Judgment by quantity. Journal of Experimental Psychology: General, 123, 1, 1994, 21-32.

[10] Kelly, S. D., Barr, D. J., Church, R. B. and Lynch, K. Offering a Hand to Pragmatic Understanding: The Role of Speech and Gesture in Comprehension and Memory. Journal of Memory and Language, 40, 1999, 577-592.

[11] LG.PHILIPS LCD Co. URL: http://www.tinyurl.com/2hou5e/ [Last accessed on June 09, 2007].

[12] McGurk, H. and MacDonald, J. Hearing lips and seeing voices. Nature, 264, 5588, 1976, 746-748.

[13] O'Hara, K., Black, A. and Lipson, M. Everyday practices with mobile video telephony. Proceedings of the SIGCHI conference on Human Factors in computing systems, 2006.

[14] Risberg, A. and Lubker, J. Prosody and speechreading. Speech Transmission Laboratory

Quarterly Progress Report and Status Report, 4, 1978, 1-16.
[15] Robert-Ribes, J., Schwartz, J. L., Lallouache, T. and Escudier, P. Complementarity and synergy in bimodal speech: auditory, visual, and audio-visual identification of French oral vowels in noise. The Journal of the Acoustical Society of America, 103, 6, 1998, 3677-3689.
[16] Sumby, W. H. and Pollack, I. Visual Contribution to Speech Intelligibility in Noise. The Journal Of The Acoustical Society Of America, 26, 2, 1954, 212-215.
[17] Tan, D. S. Exploiting the cognitive and social benefits of physically large displays. Carnegie Mellon University, Pittsburgh, Pennsylvania, 2004.
[18] Vitkovitch, M. and Barber, P. Visible speech as a function of image quality: effects of display parameters on lipreading ability. Applied cognitive psychology, 10, 2, 1996, 121.
[19] Vroomen, J. and De Gelder, B. Perceptual effects of cross-modal stimulation: The cases of ventriloquism and the freezing phenomenon. MIT Press, Cambridge, Massachusetts (USA), 2004.