BEHAVIOUR & INFORMATION TECHNOLOGY
2018, VOL. 37, NO. 9, 874–883
https://doi.org/10.1080/0144929X.2018.1505954

Taylor & Francis
Taylor & Francis Group

Enhancing the quality of service of mobile video technology by increasing multimodal synergy

F. van der Sluis[a], E. L. van den Broek[b], A. van Drunen[c] and J. G. Beerends[d]

[a]Department of Information Studies, University of Copenhagen, Copenhagen, Denmark; [b]Department of Information and Computing Sciences, Utrecht University, Utrecht, Netherlands; [c]Amsterdam University of Applied Sciences, Amsterdam, Netherlands; [d]TNO, The Hague, Netherlands

ABSTRACT

Bandwidth is still a limiting factor for the Quality of Service (QoS) of mobile communication applications. In particular, for Voice over IP the QoS is not yet as good as for common, well-engineered, public-switched telephone networks. Multisensory communication has been identified as a possibility to moderate this limitation. One of the strengths of mobile video technology lies in its combination of visual and auditory modalities. However, one of the most salient features of mobile video applications is its small screen size. To test the potential of multimodal synergy for mobile devices, we assessed to what extent small screens affect multimodal synergy. This potential was assessed in an experiment with 54 participants, who conducted a standardised video-listening test for three talking-heads videos with a signal-to-noise ratio of –9 dB. The videos were presented on three different screen sizes, whilst keeping the video and auditory signals equal. Compared to a ground truth based on 359 participants, intelligibility was found to be significantly higher when using a large screen than when using a small screen. This indicates that mobile video technology has the potential for a significant multimodal synergy to which screen size is a substantial constraint. To optimally benefit from their multimodal potential, we offer suggestions on how to increase the effective screen size for small screen (e.g. mobile) devices and applications through elaborating the most relevant (visual) features. We conclude that knowledge about human sensory processing can alleviate the identified constraint and maximise the potential QoS of mobile video technology.

ARTICLE HISTORY
Received 4 July 2018
Accepted 16 July 2018

KEYWORDS
Multimodal; multisensory; mobile; screen size; quality of service (QoS); voice over IP (VoIP); field of view (FOV); signal-to-noise ratio (SNR)

1. Introduction

Mobile technology such as e-readers, laptops, smartphones, smartwatches and other wearables is increasingly commonplace (Johnson and Grainge 2015; Škařupová, Ólafsson, and Blinka 2016; Powell 2017; Shoukry and Gobel 2017; van den Broek 2017). Consequently, applications such as mobile video telephony, mobile television, mobile Internet, navigation and mobile games are growing in use (Bernhaupt and Pirker 2014; Hess, Knoche, and Wulf 2014; Lim et al. 2015; Shaheen, Cohen, and Martin 2017). As we will illustrate next, even more than with other multimedia, the decisive benefits of mobile video technology are strongly connected to its multimodal nature (Yuan, Ghinea, and Muntean 2015; Shaked and Winter 2016):

- First, for mobile video telephony (e.g. via a smartphone or smartwatch), its main uses reflect the decisive benefits of the technology: functional talk (22%), showing objects (28%), and social and emotional small talk (50%) (O'Hara, Black, and Lipson 2006; Jang, Kim, and Ko 2017). In comparison to audio-only telephony, contact is facilitated and the social and emotional aspects of communication are enhanced (de Gelder and Vroomen 2000). However, mobile video telephony is still used at a limited scale, this despite the potential benefits and its availability to many consumers (Baraković and Skorin-Kapov 2015).
- Second, mobile television is rapidly becoming the next high-growth consumer technology (Jung, Perez-Mira, and Wiley-Patton 2009; Hess, Knoche, and Wulf 2014). Its multimodality enables a user to be immersed whilst being on the way and privatising the user from its surroundings (Powell 2017; Rimell, Mansfield, and Hands 2007). It is this immersion which is regarded as a key benefit of the technology (Jung, Perez-Mira, and Wiley-Patton 2009).
- Third, mobile games serve a similar experience (Thompson, Nordin, and Cairns 2012; Shoukry and Gobel 2017). The mobile gaming platform is commonly used to immerse oneself away from the surroundings; for example, in China, South Korea, and Japan, mobile gaming is immensely popular,

CONTACT F. van der Sluis f.vandersluis@acm.org

especially in the context of mass transit commuting (Liu and Li 2011).

Despite the potential benefits, the available bandwidth is still a limiting factor for the Quality of Service (QoS) of modern communication applications, testified by the ongoing research on techniques that improve the QoS within a restricted and less reliable network. Example target areas are, at a terminal level, the use of different codecs; at a network level, controlling packet loss and delays; and at a user level, influencing the mean opinion scores of listening, conversations, and underlying subjective quality factors such as distortion, loudness, delay, and echo (Takahashi, Yoshino, and Kitawaki 2004). In particular for Voice over IP (VoIP), the QoS is not yet as good as common, well-engineered, Public Switched Telephone Network (PSTN) (Karapantazis and Pavlidou 2009).

Multimodal communication has already been identified as an important possibility for improving the QoS of multimedia, where a synergy between the modalities can solve part of the issues in QoS (Tasaka and Ishibashi 2002; Schulte, Chen, and Nahrstedt 2014). Utilising the multimodal nature of human perception, this paper will test if multimedia QoS can be improved without the need for extra bandwidth.

1.1. Multimodal perception in theory

Multimodal perception has several benefits over unimodal perception, indicating its potential for multimedia QoS. Sumby and Pollack (1954) were among the first to describe the synergy of our auditory and visual percepts: combined auditory-visual perception is superior to perception through either audio or vision alone (Erber 1975; Calvert, Spence, and Stein 2004; Ernst and Bülthoff 2004; Stein 2012). This synergy is especially salient in noisy surroundings, where the bimodal advantage can become as large as a 39% increase on intelligibility (Risberg and Lubker 1978).

Multimodal synergy not only enhances the intelligibility of a message presented visually as well as auditory, but also enhances aspects such as memory and emotion (de Gelder and Vroomen 2000; Janssen et al. 2013). These effects are particularly beneficial when the presented is complex (van der Sluis et al. 2014) or the user is under high cognitive load (Cao et al. 2010). For example, adding visual gestures to auditory speech improves the quality of the memory for speech (Kelly et al. 1999) and both modalities supply complementary information about emotions (de Gelder and Vroomen 2000); both the face and voice are effectively combined (Freeman and Ambady 2011). The latter explains the main use of (mobile) video telephony for social and emotional talk (O'Hara, Black, and Lipson 2006). More generally, it illustrates the superiority of multimodal communication over unimodal communication (cf. Ernst and Bülthoff 2004).

Synergy is regarded as a primary evaluation criterion for the usability of multimodal interfaces (Perakakis and Potamianos 2008). However, these bimodal advantages can benefit from or be restricted by three types of concerns.

(1) Auditory concerns such as noise need to be taken into account (Watts 2008).
(2) Visual concerns are of influence (Fernandez-Lopez, Martinez, and Sukno 2017); for example, temporal frequency, spatial resolution (Calvert, Spence, and Stein 2004, Chapter 11; Stein 2012), and noise (Rimell, Mansfield, and Hands 2007). Both auditory and visual concerns relate to the available bandwidth and characteristics of the device.
(3) Bimodal concerns need to be considered: (a) spatial coherence (or ventriloquism effect): the perceived direction of an auditory stimulus is altered due to the influence of a visual stimulus (Vroomen et al. 2004); (b) source coherence: two sources behave in a way that an (auditory) stimulus is ascribed to them both (McGurk and MacDonald 1976); (c) temporal coherence: different events take place at (almost) the same time and are thus seen as one stimulus (Dixon and Spitz 1980); and (d) visual dominance: visual stimuli tend to prevail over non-visual stimuli. Even for speech perception, which is generally considered to be an auditory function, vision can strongly alter the quality of the auditory percept (Ferris and Sarter 2008; Shams and Kim 2010).

1.2. Multimodal perception in mobile practice

Of the possible restrictions to bimodal synergy, the influence of screen size has received little attention within the context of mobile devices. Their small screen is, however, one of their most salient features and screen size is likely to influence the bimodal advantages and, thus, the user experience. Jung, Perez-Mira, and Wiley-Patton (2009) plead for more research on the influence of screen size, concluding that: 'a small screen (..) is considered the fatal disadvantage of mobile TV service' (129) (cf. Yuen, Tang, and Wang 2002).

Research on screen size and multimodal synergy indicates a potential to improve the QoS of mobile video technology (Tan et al. 2006; Kim 2017). Findings on

the influence of primarily large screens support this plea, as larger screens have been found to influence variables such as arousal, sense of presence, attention and memory, connectedness, and game immersion (Grabe et al. 1999; Thompson, Nordin, and Cairns 2012; Kim 2017). For most of these variables, the effects can be summarised as intensifying the values. Hence, 'the larger, the better' seems to hold. Within reasonable limits, spatial resolution or information throughput has been shown less important for bimodal synergy (Frowein et al. 1991). These findings on bimodal synergy suggest that optimal bandwidth utilisation might not be the only critical factor to consider for the QoS of mobile video technology.

Combining the findings on the effects of large screens as well as information throughput, this study examines whether multimodal synergy can be improved without the need for extra bandwidth capacity. This is examined in an experiment through increasing the Field of View (FOV) (i.e. perceived size) of talking-heads video material, whilst keeping the amount of information constant. The experiment is performed above basic levels of visual acuity (i.e. one arc minute), to assure that participants do not acquire extra information and that any effect can be ascribed to enlarging multimodal synergy. Expected is that an increase in FOV enhances the bimodal advantages. To answer this hypothesis, the intelligibility of a message presented auditory as well as visually is measured. The relative importance of the visual compared to the auditory modality is increased by adding noise to the auditory channel. Consequently, changes in the visual channel are expected to have a greater effect on the intelligibility.

In the next section, we present the research methods, including information on the participants, material, apparatus, and procedure as well as specifications on how the FOV and Signal-to-Noise Ratio (SNR) are determined. Section 3 presents the results of the experiment. Subsequently, Section 4 presents a discussion of the results and its implications. We close with a brief conclusion in Section 5.

2. Method

To study the influence of FOV, a within-subjects design was used evaluating the effects of screen size (i.e. small, medium, and large), video (i.e. 1, 2, and 3), and their sequence (i.e. first, second, and third). The design served to counterbalance any order effects. This gave a total of 36 ($3! \times 3!$) different conditions, based on the possible number of combinations of three screen sizes and three videos. The order of participation determined to which condition a subject was assigned. For example, the first subject was assigned to the first condition.

2.1. Participants

Fifty-four subjects (mean age: 20.3; range: 18–28) participated in the research. The participants were recruited at the university campus. They participated either on a voluntary basis or received study credits for their participation in the experiment. All participants had a (corrected to) normal vision and hearing. 96.2% of the participants judged their level of English as either good or reasonable.

2.2. Material: audio listening test

Participant's English level was verified using a standardised audio-only listening test of the Dutch Central Institute for Testing (CITO 2018). The test was part of the English listening exams of secondary, pre-university, education in the Netherlands. It consisted of 12 parts that lasted in total about 10 minutes and that depicted an interview with a probation officer about his job. After each part, participants' comprehension was evaluated using a multiple-choice question and scored using norm-based correction forms. For all subjects, their English level was found to be sufficient ($mean = 8.89$, $SD = 1.98$, the maximum score possible is 12).

2.3. Material: video listening test

To evaluate the intelligibility of a message presented auditory as well as visually, a set of three videos were used from a standardised video listening test (CITO 2018). The video listening test depicted an interview with an exchange student. The videos were selected such that the face of the person talking was visible most of the time with the camera focused on the face; that is, talking-heads material, shown to be especially beneficial from bimodal presentation (Rimell, Mansfield, and Hands 2007). The video durations were 1′09″, 1′16″, and 1′17″.

The intelligibility of the videos was evaluated using questions from the standardised video listening test (CITO 2018). From the original set of questions, only those questions were selected that corresponded to a part of the video in which the speaker was visible. The resulting set contained two English three-choice questions per video. The test results were evaluated using the original CITO (2018) scoring forms and scored on a [0, 1] scale per video. No restrictions were made on answering time, though the answering times were measured as proxy of intelligibility (van der Sluis,

Ginn, and van der Zee 2016). The answering times were summed over both questions per video.

2.4. Apparatus

A computer with a 17-inch flat CRT screen and a headset was used to administer the audio-only and video listening tests. The subjects had to keep their heads between a square of ropes surrounding it at forehead height, securing a fixed distance of 80 *cm* to the screen. In addition, the chair and keyboard were also placed at a fixed position. The experimental setup is shown in Figure 1, where the rope construction is highlighted by dashed lines.

Three screen sizes were used, as specified in Table 1. All screen sizes were displayed on the same screen with a screen resolution of 1024×768. E-prime 2.0 was used as experiment platform, including for presenting the stimuli and measuring the per-video intelligibility test and answering times.

All videos had a resolution of 178×142 pixels, the resolution of the smallest screen size. To keep the amount of information constant, the smallest spatial resolution of 178×142 pixels was upscaled to the required spatial resolution. Hence, no extra information was given through the visual channel. The upscaling method used is the default 'high quality' algorithm of the Microsoft Windows video processing environment DirectShow (Microsoft 2018), which is an enhanced bilinear method (Srinivasan et al. 2004).

The used minimal resolution of 178×142 pixels corresponds to a dot pitch (i.e. distance between two pixels' centres) in centimetres of

$$0.034 \approx d_{cm}/d_p, \tag{1}$$

where d_p is the diagonal in pixels and d_{cm} the diagonal in centimetres. The corresponding values for d_p and d_{cm} are given in Table 1. This dot pitch is considerably above what is observable by the human eye. At a distance of $d = 80\,cm$, a person with normal visual acuity of $\alpha = 1'$ (one arc minute or $1^\circ/60$) is capable of seeing a minimal dot pitch of

$$0.023 = 80 \tan\left(\alpha \times \frac{\pi}{180}\right), \tag{2}$$

as is specified in Westheimer (1979). In Pixels Per Centimeter (PPC), this corresponds to 42.97 PPC spatial resolution for the human eye at 80 *cm* distance, compared to 29.64 PPC used in this experiment.

Figure 1. Experimental setting. The dashed lines indicate the rope construction used to secure a fixed distance between the participant and the screen.

2.5. Determination of field-of-view (FOV)

As an indicator of FOV, the Instantaneous Field of View (IFOV) was calculated for the azimuth (horizontal) direction. The IFOV combines the FOV of both eyes under a fixed head position. It is defined as following:

$$IFOV_{\text{azimuth}} = 2 \tan^{-1}\left(\frac{D_c + d_e}{2l}\right), \tag{3}$$

where D_c is the diameter of the screen, d_e is the eye separation parameter (Banbury 1983), and l is the distance from the eyes to the screen.

As eye separation parameter $d_e = 0.63\,\text{cm}$ is used; that is, an empirically supported approximation of the mean inter-pupillary distance for adults (Dodgson 2004). The diameter is approximated by the diagonal. Table 1 provides the IFOV and length of the diagonal for each of the screen sizes used.

2.6. Determination of signal-to-noise ratio (SNR)

Previous studies have shown that multimodal synergy increases when unimodal signal quality decreases (Erber 1975). For all videos, a SNR of −9 dB was used to enhance synergy to utilise this effect. This SNR was chosen in support of external validity. Speech elements below the noise level still contribute to the intelligibility (Drullman 1995) and compared to the SNR range of −6 to −30 dB used in other studies (e.g. Erber 1975), the SNR has been kept low. The expected increase in synergy was confirmed by a pilot study (N=6).

The SNR was computed as follows:

$$SNR = RMS_{\text{signal}} - RMS_{\text{noise}}, \tag{4}$$

Table 1. Screen size (diagonal), resolution, and instantaneous field of view (IFOV) of each screen used in the experiment.

Diagonal (cm)	Resolution (pixels)	Diagonal (pixels)	IFOV (degree)
43.18	1024×768	1280.00	30.63
17.58	394×314	521.07	12.99
7.68	178×142	227.70	5.95

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

$$RMS = 20\log_{10}\frac{X}{Xr}, \quad (5)$$

where X is either the power of the noise or signal and Xr is the power of the reference point of the used dB scale. For all dB values, dB relative to full scale (dBFS) was used as unit of measurement for amplitude levels, with as reference point the digital system's maximum output level. As noise source, white noise was used: a random signal, which adds an equal amount of energy across all frequencies. The RMS for both the noise and signal were calculated and normalised using Syntrillium Software Corporation's Cool Edit Pro 2.1. Normalisation was realised by taking the signal's peak amplitude and amplify the entire signal with a scalar such that its RMS reaches the predetermined level, which is possible without clipping. Consequently, for all audio signals that have the same loudness level was secured.

2.7. Procedure

The subjects were told that they were conducting a listening and a video-listening test for which they should remember as much as possible from the video. Furthermore, they were told to sit still and keep their head stable. They were informed about a video camera the experiment leader used to inspect the proper participation in the experiment.

The experiment consisted of the following four phases:

(1) Some questions concerning general demographic data were asked, namely: name, sex, age, occupation, and nationality.
(2) The audio-only English listening test was assessed (CITO 2018).
(3) Three videos in three different screen sizes were shown in one of the 36 possible orders. Each video was followed by two multiple-choice questions to test for intelligibility.
(4) Lastly, some questions were asked concerning the experience with the experiment.

The total duration of the experiment was approximately 30 minutes.

3. Results

As expected, the SNR reduced the intelligibility of the standardised CITO video listening test. A one-tailed t-test showed a significant difference between the average norm results per question of the CITO (2018) (N=359; M=0.85, $SD = 0.11$) and the current accuracy scores for the large screen size ($M = 0.74$, $SD = 0.16$); $t(53) = -2.80$, $p = .007$, $\eta_p^2 = 0.129$. With a reduction of 12.94%, this shows an overall modest influence of the added noise.

The descriptive statistics of accuracy scores and answering times per screen size, video, and sequence are shown in Table 2. The accuracy scores are on a scale of 0–1, where 1 means all questions were answered correctly. The answering times represent the total time spent to answer both questions per video. The effect of screen size, video, and sequence on accuracy score and answering time were analysed using a Multivariate Analysis of Variance (MANOVA), shown in Table 3.

The MANOVA of accuracy score by screen size showed that screen size significantly influences intelligibility ($F(2, 135) = 4.60$, $p = .012$, $\eta_p^2 = .064$). Furthermore, the correlation between screen size and accuracy score was $r(160) = .213$, $p = .007$. Figure 2 illustrates this relation, using Cousineau (2005)'s confidence intervals to make the effect size graphically visible.

Post-hoc Bonferonni comparisons for the effects of screen size on accuracy score revealed no significant results on the comparison between small (s) and medium (m) ($\Delta(s, m) = 0.07$, $SE = 0.06$, p=.734) and medium (m) and large (l) ($\Delta(m, l) = 0.10$, $SE = 0.06$, p=.205). The difference between and small (s) and large (l) was significant ($\Delta(s, l) = 0.17$, $SE = 0.06$, p=.009).

For answering time, no effects were found for screen size ($F(2, 135) = 0.10$, $p = .908$, $\eta_p^2 = .001$), which clearly shows from the data as well (see Table 2). Nonetheless, answering time did correlate with accuracy score ($r(160) = -.198$, $p = .011$), confirming its value as a proxy of intelligibility.

The three videos differed significantly in difficulty, as was also revealed by the MANOVA for both accuracy

Table 2. Accuracy scores and answering times with 5% confidence intervals per screen size, video, and sequence.

		Accuracy scores			Answering times		
Screen size	IFOV	Mean	CI – LB	CI – UB	Mean	CI – LB	CI – UB
Small	5.95°	.57	.49	.66	30.33	27.13	33.54
Medium	12.99°	.64	.55	.73	29.53	25.71	33.36
Large	30.63°	.74	.66	.82	30.46	26.85	34.06
Video							
1		.53	.43	.62	32.76	29.42	36.10
2		.58	.51	.66	36.05	32.89	39.22
3		.84	.78	.91	21.51	18.67	24.33
Sequence							
First		.66	.56	.75	30.14	26.63	33.66
Second		.64	.55	.73	30.80	27.08	34.53
Third		.66	.57	.74	29.37	25.96	32.79
Mean (SD)		0.65 (0.32)			30.11 (12.95)		

Note: CI stands for confidence interval, LB for lower bound, and UB for upper bound.

Table 3. Multivariate analysis of variance for accuracy scores and answering times by screen size, video, and sequence.

	Accuracy scores				Answering times			
Source	SS	df	MS	*F*	SS	df	MS	*F*
Screen (S)	3.05	2	1.53	4.60*	27.22	2	13.61	0.10
Video (V)	12.20	2	6.10	18.36***	6277.27	2	3138.64	22.27***
Sequence (SE)	0.05	2	0.03	0.07	55.29	2	27.65	0.20
S × V	0.10	4	0.03	0.07	485.93	4	121.48	0.86
S × SE	0.91	4	0.23	0.69	31.52	4	7.88	0.06
V × SE	1.32	4	0.33	0.99	484.55	4	121.14	0.86
S × V × SE	3.72	8	0.47	1.40	603.24	8	75.41	0.54
Error	44.83	135	0.33		19027.8	135	140.95	

Note: Explained variance by MANOVA for accuracy score $R^2 = .323$*** and for answering time $R^2 = .295$**. *$p < .05$, **$p < .01$, ***$p < .001$

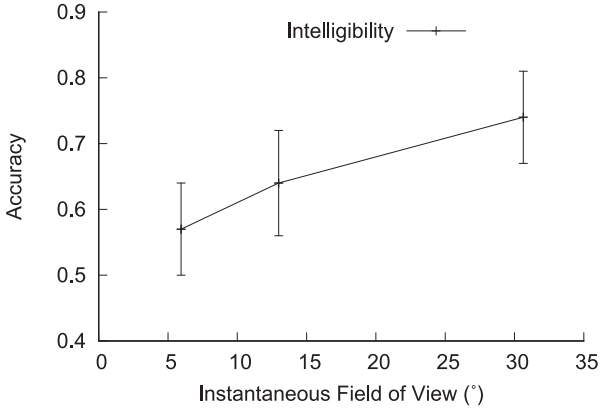

Figure 2. Relation between Field of View (FOV) and Accuracy, showing that a higher FOV leads to an increase in intelligibility. The vertical bars depict confidence intervals excluding between-subject variability (cf. Cousineau 2005).

score ($F(2, 135) = 18.36$, $p<.001$, $\eta_p^2 = .214$) and answering time ($F(2, 135) = 22.27$, $p<.001$, $\eta_p^2 = .248$). The influence of sequence was non-significant for both accuracy score ($F(2, 135) = 0.74$, $p = .928$, $\eta_p^2 = .001$) and answering time ($F(2, 135) = 0.20$, $p = .822$, $\eta_p^2 = .003$), indicating that there was no learning effect within the different trials that each subject performed. Furthermore, the English level as tested with the standardised audio listening test neither correlated with accuracy score on the video test ($r(52) = .083$, $p = .550$) nor with answering time ($r(52) = -.016$, $p = .911$). This indicates that differences in the level of English did not influence intelligibility and, instead, that differences in intelligibility can be wholly attributed to the influence of screen size and video. Finally, gender did not influence the correlation between screen size and accuracy. For both men and women the correlation remained r=.21.

4. Discussion

In line with Sumby and Pollack (1954), 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 standardised video-listening test for three different screen sizes, indicating that screen size is indeed an influential factor in intelligibility. A smaller screen size accounted for as much as a 22.30% reduction in intelligibility. Hence, through utilising standardised intelligibility tests, this research specifies and quantifies the fundamental constraint that small screens place on the advantages of bimodal perception.

The effect of screen size appears to be robust, showing a quite consistent and gradual increase in synergy with an increase in screen size (see Figure 2). Thus, even when it is possible to reduce the distance to the screen, the synergy is still likely to benefit from a larger screen size. Findings on the positive effects of large screens on other psychological variables further support this expectation (Grabe et al. 1999; Kim 2017). This result does ask for further research, as to find the threshold above which an increase in FOV does not further increase bimodal synergy.

Several factors are expected to interact with the influence of screen size on bimodal synergy. The effects can be:

- Different for natural noise. The amount of noise in the auditory channel increases bimodal synergy (Sumby and Pollack 1954) which, by extension, means that noise strengthens the effects found for screen size. In this experiment, the added white noise reduced intelligibility by 12.94%. In real usage of mobile video technology, noise is likely to occur frequently and intensively (Watts 2008) but in other forms than 'clean' white noise.
- Different with other types of video material. In a realistic setting, the quality of the visual channel is likely to be less than as used in the experiment. For example, the talking face will be less prominent, making the relevant FOV smaller and the spatial resolution less.

- Different at other cognitive levels (e.g. emotional). The strength of the effect and the threshold till which any effect is salient will likely be different for variables such as emotional connectedness (van den Broek, van der Sluis, and Schouten 2010; van den Broek 2011; Janssen et al. 2013), one of the key uses of mobile video telephony (O'Hara, Black, and Lipson 2006) and possibly of mobile television and mobile games.
- Different for other applications. The type of application is likely to put a focus on other features of a channel or even other channels. For example, for non-verbal communication, a visual dominance exists (Ernst and Bülthoff 2004; Ferris and Sarter 2008). Or, when aiming for an immersive gaming experience, the possible synergy of adding audio to the predominantly visual mobile games has been proposed (Ekman et al. 2005).

These factors call for further research to the conditions under which the discovered constraint to bimodal synergy is the most salient.

Several other findings are noteworthy as well. Namely, contrary to accuracy score, answering time did not respond to screen size whereas it did respond to video difficulty. This indicates that answering time and accuracy score reflect distinct aspects of intelligibility (cf. van der Sluis, van der Zee, and Ginn 2017). It suggests a different level of processing for the contents of the video and the integration of the different modalities. In addition, the effects found appeared to be unrelated to differences in English level between participants. Given that all participants were at least reasonably capable of understanding English, they were also able to understand the spoken contents of the videos. For this sample of participants, the used experimental setting was sensitive to differences in intelligibility rather than to differences in comprehensibility. Lastly, contrary to other studies on the effects of large screens on viewer experiences (e.g. Grabe et al. 1999), no gender differences were found on the importance of screen size.

This study adds to a body of research that predominantly shows the effects of big screens on a range of psychological variables. It specifies and quantifies the constraint that small screens place by showing the effects of screen size on multimodal synergy. It is this constraint on bimodal synergy that likely underlies a range of other variables known to be affected by larger screens, such as arousal, sense of presence, attention and memory, connectedness, and even game immersion (Grabe et al. 1999; Thompson, Nordin, and Cairns 2012; Kim 2017). As such, multimodal synergy offers a theoretical framework to study and explain the importance of mobile screen size for various variables related to the mobile user experience.

This study points to a fundamental but commonplace constraint on basic human multimodal perception and places it in the context of the field of mobile video technology. The constraint of screen size to bimodal synergy is important if mobile video applications are to benefit from their bimodal nature. In a setting where through a SNR of −9 dB the influence of the visual stimuli was enlarged, a lower screen size already reduced intelligibility by 22.30% compared to the large screen size. Hence, having a profound effect on bimodal advantages and, thus, on the added value of mobile video technology. Moreover, the benefits of bimodal perception are beyond mere intelligibility alone, on memory and emotion as well (Kelly et al. 1999; Roring, Hines, and Charness 2006), making it a key factor in mobile user experience.

The identified constraint shows both the vulnerability and strength of mobile video technology: When the constraints are met, the mobile user experience can fully benefit from the potential bimodal advantages. It shows one of the possible reasons for the absence of large scale success of mobile video telephony and supplied evidence for one of the possible threats to mobile television. But it also shows the potential of improving the QoS of mobile video technology. Multimodal synergy has the potential to alleviate auditory and visual issues that emerged in parallel with mobile technology (Harper, Yesilada, and Chen 2011; Škařupová, Ólafsson, and Blinka 2016; van den Broek 2017), whilst using the same bandwidth.

In a search for a solution to possible decreased levels of synergy, the most straightforward option would be to increase the screen size. This can be done by, for example, lightweight high-resolution video glasses (Costanza et al. 2006) and flexible electronic paper (Rogers, Someya, and Huang 2010). However, as this study has also shown, fundamental characteristics of human perception underlie the bimodal advantages. These characteristics create a less straightforward but attractive solution, by allowing to increase bimodal synergy without increasing the screen size. For example, when aiming at intelligibility, the lips cause the largest part of the bimodal synergy (Vroomen et al. 2004). Hence, the lips can be made the core of the visual channel (Ouni and Gris 2018). Instead, when aiming at emotional connectedness, the whole face becomes an important information holder (de Gelder and Vroomen 2000). Removing peripheral information from the visual channel is likely to enhance bimodal synergy for emotional connectedness. This points to both a technical solution, by zooming in on the features salient to bimodal synergy, and to a content solution, by using cinematic techniques allowing for a high zoom level.

5. Conclusion

More than half a century ago, Sumby and Pollack (1954) described the synergy of our auditory and visual percepts; seeing someone speak helps hearing what he says. Par excellence, it illustrates the holistic process underlying human multisensory perception. The current study places the work of Sumby and Pollack (1954) on human perception and information processing in the context of mobile video technology.

The presented study revealed an influential factor to the success of mobile video technology: the limited synergy of audio and video with small screens. It showed one of the possible reasons for the absence of a large scale success of mobile video telephony and supplied evidence for one of the possible threats to mobile video applications. Whilst at the same time, this study illustrated how mobile video applications can benefit from adapting to fundamental characteristics of human multimodal perception. In particular, with small screens, we advice to remove peripheral information from the visual channel in order to enhance bimodal synergy. Our findings point to some new and unexpected directions for future research on improvement in QoS of mobile video telephony. It shows that the effects of audio and video quality cannot be treated separately and that any improvement in QoS depends on their synergetic effects.

Acknowledgments

The CITO (2018), 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 for their cooperation and fruitful discussions during this study. Last, we thank all subjects for their participation in this study.

Disclosure statement

No potential conflict of interest was reported by the authors.

Funding

The Dutch organisation for scientific research (NWO) is gratefully acknowledged for funding the IPPSI-KIEM project Adaptive Text-Mining (ATM) (project-number: 628.005.006), in which Frans van der Sluis and Egon L. van den Broek cooperate.

References

Microsoft. 2018. "About DirectShow." Accessed March 16, 2018. https://msdn.microsoft.com/en-us/library/windows/desktop/dd373389(v=vs.85).aspx.

CITO. 2018. Accessed June 15, 2018. https://msdn.microsoft.com/enus/library/windows/desktop/dd373389(v=vs.85).aspx.

Banbury, J. R. 1983. "Wide Field of View Head-up Displays." *Displays* 4 (2): 89–96.

Baraković, S., and L. Skorin-Kapov. 2015. "Multidimensional Modelling of Quality of Experience for Mobile Web Browsing." *Computers in Human Behavior* 48: 314–332.

Bernhaupt, R., and M. M. Pirker. 2014. "User Interface Guidelines for the Control of Interactive Television Systems via Smart Phone Applications." *Behaviour & Information Technology* 33 (8): 784–799.

Calvert, G. A., C. Spence, and B. E. Stein. 2004. *The Handbook of Multisensory Processes*. Cambridge: The MIT Press / A Bradford Book.

Cao, Y., F. van der Sluis, M. Theune, R. op den Akker, and A. Nijholt. 2010. "Evaluating Informative Auditory and Tactile Cues for In-vehicle Information Systems." In *Proceedings of the 2nd International Conference on Automotive User Interfaces and Interactive Vehicular Applications (AutomotiveUI'10)*, edited by A. K. Dey, A. Schmidt, S. Boll, and A. L. Kun, Pittsburgh, PA, 102–109, November 11–12. New York: ACM.

Costanza, E., S. A. Inverso, E. Pavlov, R. Allen, and P. Maes. 2006. "eye-q: Eyeglass Peripheral Display for Subtle Intimate Notifications." In *MobileHCI '06: ACM Proceedings of the 8th Conference on Human-computer Interaction with Mobile Devices and Services*, edited by Mika Röykkee, A. Kaikkonen, M. Nieminen, and K. Väänänen-Vainio Mattila, Vol. 159 of *ACM International Conference Proceeding Series*, Espoo, Finland, 211–218, September 12–15. New York: ACM Press.

Cousineau, D. 2005. "Confidence Intervals in Within-subject Designs: A Simpler Solution to Loftus and Masson's Method." *Tutorials in Quantitative Methods for Psychology* 1 (1): 42– 45.

de Gelder, B., and J. Vroomen. 2000. "The Perception of Emotions by Ear and by Eye." *Cognition and Emotion* 14 (3): 289–311.

Dixon, N. F., and L. Spitz. 1980. "The Detection of Auditory Visual Desynchrony." *Perception* 9 (6): 719–721.

Dodgson, N. A. 2004. "Variation and Extrema of Human Interpupillary Distance." *Proceedings of SPIE (Stereoscopic Displays and Virtual Reality Systems)* 5291: 36–46.

Drullman, R. 1995. "Speech Intelligibility in Noise: Relative Contribution of Speech Elements Above and Below the Noise Level." *The Journal of the Acoustical Society of America* 98 (3): 1796–1798.

Ekman, I., L. Ermi, J. Lahti, J. Nummela, P. Lankoski, and F. Mäyrä. 2005. "Designing Sound for a Pervasive Mobile Game." In *Proceedings of the 2005 ACM SIGCHI International Conference on Advances in Computer Entertainment Technology (ACE 2005)*, edited by N. Lee, Valencia, Spain, 110–116, June 15–17. New York: ACM.

Erber, N. P. 1975. "Auditory-visual Perception of Speech." *Journal of Speech and Hearing Disorders* 40 (4): 481–492.

Ernst, M. O., and H. H. Bülthoff. 2004. "Merging the Senses Into a Robust Percept." *TRENDS in Cognitive Sciences* 8 (4): 162–169.

Fernandez-Lopez, A., O. Martinez, and F. M. Sukno. 2017. "Towards Estimating the Upper Bound of Visual-speech Recognition: The Visual Lip-reading Feasibility Database."

In *Proceedings of the 12th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2017)*, edited by K. Bowyer, R. Chellappa, J. Cohn, H. Gunes, A. O'Toole, C. Pelachaud, Y. Tong, and V. Patel, Washington, DC, 208–215, May 30–June 03. Los Alamitos, CA: IEEE Computer Society.

Ferris, T. K., and N. B. Sarter. 2008. "Cross-modal Links Among Vision, Audition, and Touch in Complex Environments." *Human Factors* 50 (1): 17–26.

Freeman, J. B., and N. Ambady. 2011. "When two Become One: Temporally Dynamic Integration of the Face and Voice." *Journal of Experimental Social Psychology* 47 (1): 259–263.

Frowein, H. W., G. F. Smoorenburg, L. Pyters, and D. Schinkel. 1991. "Improved Speech Recognition Through Videotelephony: Experiments with the Hard of Hearing." *IEEE Journal on Selected Areas in Communications* 9 (4): 611–616.

Grabe, M. E., M. Lombard, R. D. Reich, C. C. Bracken, and T. B. Ditton. 1999. "The Role of Screen Size in Viewer Experiences of Media Content." *Visual Communication Quarterly* 6 (2): 4–9.

Harper, S., Y. Yesilada, and T. Chen. 2011. "Mobile Device Impairment … Similar Problems, Similar Solutions?." *Behaviour & Information Technology* 30 (5): 673–690.

Hess, J., H. Knoche, and V. Wulf. 2014. "Thinking Beyond the Box: Designing Interactive TV Across Different Devices." *Behaviour & Information Technology* 33 (8): 781–783.

Jang, S. B., Y. G. Kim, and Y.-W. Ko. 2017. "Mobile Video Communication based on Augmented Reality." *Multimedia Tools and Applications* 76 (16): 16893–16909.

Janssen, J. H., P. Tacken, J. J. G. de Vries, E. L. van den Broek, J. H. D. M. Westerink, P. Haselager, and W. A. IJstelstein. 2013. "Machine Beats Human Emotion Recognition through Audio, Visual, and Physiological Modalities." *Human Computer Interaction* 28 (6): 479–517.

Johnson, C., and P. Grainge. 2015. *Promotional Screen Industries*. Oxon: Routledge / Taylor & Francis Group.

Jung, Y., B. Perez-Mira, and S. Wiley-Patton. 2009. "Consumer Adoption of Mobile TV: Examining Psychological Flow and Media Content." *Computers in Human Behavior* 25 (1): 123–129.

Karapantazis, S., and F.-N. Pavlidou. 2009. "VoIP: A Comprehensive Survey on a Promising Technology." *Computer Networks* 53 (12): 2050–2090.

Kelly, S. D., D. J. Barr, R. B. Church, and K. Lynch. 1999. "Offering a Hand to Pragmatic Understanding: The Role of Speech and Gesture in Comprehension and Memory." *Journal of Memory and Language* 40 (4): 577–592.

Kim, K. J. 2017. "Shape and Size Matter for Smartwatches: Effects of Screen Shape, Screen Size, and Presentation Mode in Wearable Communication." *Journal of Computer-Mediated Communication* 22 (3): 124–140.

Lim, J. S., S. Y. Ri, B. D. Egan, and F. A. Biocca. 2015. "The Cross-platform Synergies of Digital Video Advertising: Implications for Cross-media Campaigns in Television, Internet and Mobile TV." *Computers in Human Behavior* 48: 463–472.

Liu, Y., and H. Li. 2011. "Exploring the Impact of Use Context on Mobile Hedonic Services Adoption: An Empirical Study on Mobile Gaming in China." *Computers in Human Behavior* 27 (2): 890–898.

McGurk, H., and J. MacDonald. 1976. "Hearing Lips and Seeing Voices." *Nature* 264 (5588): 746–748.

O'Hara, K., A. Black, and M. Lipson. 2006. "Everyday Practices with Mobile Video Telephony." In *Proceedings of the SIGCHI Conference on Human Factors in computing systems*, edited by R. Grinter, T. Rodden, P. Aoki, E. Cutrell, R. Jeffries, and G. Olson, 871–880. Montreal, QC: ACM Press.

Ouni, S., and G. Gris. 2018. "Dynamic Lip Animation from a Limited Number of Control Points: Towards an Effective Audiovisual Spoken Communication." *Speech Communication* 96: 49–57.

Perakakis, M., and A. Potamianos. 2008. "Multimodal System Evaluation using Modality Efficiency and Synergy Metrics." In *Proceedings of the 10th International Conference on Multimodal Interfaces (ICMI'08)*, edited by V. Digalakis, A. Potamianos, M. Turk, R. Pieraccini, and Y. Ivanov, Chania, Crete, Greece, 9–16, October 20–22. New York: ACM.

Powell, H. 2017. "Always On: Mobile Culture and its Temporal Consequences." Chap. 4, 99–117. London: World Scientific Publishing Europe Ltd.

Rimell, A. N., N. J. Mansfield, and D. Hands. 2007. "The Influence of Content, Task and Sensory Interaction on Multimedia Quality Perception." *Ergonomics* 51 (2): 85–97.

Risberg, A., and J. Lubker. 1978. "Prosody and Speechreading." *Speech Transmission Laboratory Quarterly Progress Report and Status Report* 4: 1–16.

Rogers, J. A., T. Someya, and Y. Huang. 2010. "Materials and Mechanics for Stretchable Electronics." *Science* 327 (5973): 1603–1607.

Roring, R. W., F. G. Hines, and N. Charness. 2006. "Age-related Identification of Emotions at Different Image Sizes." *Human Factors* 48 (4): 675–681.

Schulte, S., S. Chen, and K. Nahrstedt. 2014. "Stevens' Power Law in 3D Tele-immersion: Towards Subjective Modeling of Multimodal Cyber interaction." In *Proceedings of the 22nd ACM International Conference on Multimedia*, edited by K. A. Hua, Y. Rui, R. Steinmetz, A. Hanjalic, A. (P.) Natsev, and W. Zhu, Orlando, FL, 1133–1136, November 3–7. New York: ACM Press.

Shaheen, S., A. Cohen, and E. Martin. 2017. "Smartphone App Evolution and Early Understanding from a Multimodal App User Survey." Chap. 10, 149–164. Lecture Notes in Mobility (LNMOB). Cham: Springer International Publishing AG.

Shaked, N., and U. Winter. 2016. *Design of Multimodal Mobile Interfaces*. Berlin: Walter De Gruyter.

Shams, L., and R. Kim. 2010. "Crossmodal Influences on Visual Perception." *Physics of Life Reviews* 7 (3): 269–284.

Shoukry, L., and S. Gobel. 2017. "Reasons and Responses: A Multimodal Serious Games Evaluation Framework." *IEEE Transactions on Emerging Topics in Computing*. doi:10.1109/TETC.2017.2737953.

Škařupová, K., K. Ólafsson, and L. Blinka. 2016. "The Effect of Smartphone Use on Trends in European Adolescents' Excessive Internet Use." *Behaviour & Information Technology* 35 (1): 68–74.

Srinivasan, S., P. J. Hsu, T. Holcomb, K. Mukerjee, S. L. Regunathan, B. Lin, J. Liang, M.-C. Lee, and J. Ribas-Corbera. 2004. "Windows Media Video: Overview and

Applications." *Signal Processing: Image Communication* 19 (9): 851–875.

Stein, B. E. 2012. *The New Handbook of Multisensory Processing*. Cambridge: The MIT Press.

Sumby, W. H., and I. Pollack. 1954. "Visual Contribution to Speech Intelligibility in Noise." *The Journal of the Acoustical Society of America* 26 (2): 212–215.

Takahashi, A., H. Yoshino, and N. Kitawaki. 2004. "Perceptual QoS Assessment Technologies for VoIP." *IEEE Communications Magazine* 42 (7): 28–34.

Tan, D. S., D. Gergle, P. Scupelli, and R. Pausch. 2006. "Physically Large Displays Improve Performance on Spatial Tasks." *ACM Transactions on Computer-Human Interaction* 13 (1): 71–99.

Tasaka, S., and Y. Ishibashi. 2002. "Mutually Compensatory Property of Multimedia QoS." Proceedings of 2002 IEEE International Conference on Communications, Vol. 2, New York, USA, 1105–1111.

Thompson, Matt, A. Imran Nordin, and Paul Cairns. 2012. "Effect of Touch-screen Size on Game Immersion." Proceedings of the 26th Annual BCS Interaction Specialist Group Conference on People and Computers, BCS-HCI '12, Swinton, UK, 280–285. British Computer Society.

van den Broek, E. L. 2011. "Affective Signal Processing (ASP): Unraveling the mystery of emotions." PhD diss., Human Media Interaction (HMI), Faculty of Electrical Engineering, Mathematics, and Computer Science, University of Twente, Enschede.

van den Broek, E. L. 2017. "ICT: Health's Best Friend and Worst Enemy?" In *BioSTEC 2017: 10th International Joint Conference on Biomedical Engineering Systems and Technologies, Proceedings Volume 5: HealthInf*, edited by E. L. van den Broek, A. Fred, H. Gamboa, and M. Vaz, 611–616, February 21–23. Porto: SciTePress – Science and Technology Publications, Lda.

van den Broek, E. L., F. van der Sluis, and Th. E. Schouten. 2010. "User-centered Digital Preservation of Multimedia." *ERCIM (European Research Consortium for Informatics and Mathematics) News* 80: 45–47.

van der Sluis, F., J. H. Ginn, and T. van der Zee. 2016. "Explaining Student Behavior at Scale: The Influence of Video Complexity on Student Dwelling Time." Proceedings of the Third (2016) ACM Conference on Learning @ Scale, L@S '16, New York, NY, 51–60. ACM.

van der Sluis, F., E. L. van den Broek, R. J. Glassey, E. M. A. G. van Dijk, and F. M. G. de Jong. 2014. "When Complexity Becomes Interesting." *Journal of the American Society for Information Science and Technology* 65 (7): 1478–1500.

van der Sluis, F., T. van der Zee, and J. H. Ginn. 2017. "Learning About Learning at Scale: Methodological Challenges and Recommendations." Proceedings of the Fourth (2017) ACM Conference on Learning @ Scale, L@S '17, New York, NY, 131–140. ACM.

Vroomen, J., M. Keetels, B. de Gelder, and P. Bertelson. 2004. "Recalibration of Temporal Order Perception by Exposure to Audio-visual Asynchrony." *Cognitive Brain Research* 22 (1): 32–35.

Watts, L. 2008. "Advanced Noise Reduction for Mobile Telephony." *IEEE Computer* 41 (8): 72–79.

Westheimer, G. 1979. "The Spatial Sense of the Eye. Proctor Lecture." *Investigative Ophthalmology & Visual Science* 18 (9): 893–912.

Yuan, Z., G. Ghinea, and G.-M. Muntean. 2015. "Beyond Multimedia Adaptation: Quality of Experience-aware Multi-sensorial Media Delivery." *IEEE Transactions on Multimedia* 17 (1): 104–117.

Yuen, P. C., Y. Y. Tang, and P. S. P. Wang. 2002. *Multimodal Interface for Human–machine Communication*, Series in Machine Perception and Artificial Intelligence, Vol. 48. River Edge, NJ: World Scientific Publishing Co.