

## **Bachelorthesis Psychology**

**A. van Drunen**

# **The effects of screen-size on intelligibility in multimodal communication**

**Dr. R. van Eijk**  
Supervisor on behalf of:

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



**Telematica**  
*Instituut*



**Universiteit Twente**  
*de ondernemende universiteit*

# The effects of screen-size on intelligibility in multimodal communication

A. van Drunen (s0063703)

[a.vandrunen@student.utwente.nl](mailto:a.vandrunen@student.utwente.nl)

## ABSTRACT

This study shows screen size has a profound effect on the integration of multimodal (visual and auditory) stimuli. This was measured by the intelligibility of a video-listening test for three screen sizes: mobile phone, PDA and PC monitor. This study indicated larger effects of a temporal latency of 250ms for the largest screen size. These results are used in the context of mobile video telephony. The possible benefits, restrictions and improvements are discussed for this technique based on the findings of this study.

## 1. Introduction

Person-to-person mobile video telephony is nowadays available to many consumers, due to the growing availability of 3G networks and services and the increasing abilities of new mobile phones. Despite this growing availability, the use of mobile video telephony still lacks success, while several benefits from using this type of communicative medium can be expected. These benefits are mainly in social communication. Mobile video telephony is said to create a greater connectedness between its users, enhancing the social and emotional aspects of communication (Short, Williams, & Christie, 1976). Ethnographic research agrees with this and shows that the current use of mobile video telephony is in functional talk (22%), in showing objects (28%), and primarily in social and emotional small talk (50%) (O'Hara, Black, & Lipson, 2006). On functional talk, some evidence shows that primarily for complex tasks, task efficiency is enhanced by video (Kraut, Gergle, & Fussell,

2002). For showing objects, conversational effectiveness is increased by having visual information about objects (Whittaker, 2003).

These benefits cannot be seen without a comparison to the accompanying costs. These costs are, besides the (decreasing) financial costs, the difficulties of using mobile video telephony. These difficulties can be explained in social and practical barriers of use; e.g. privacy issues (O'Hara et al., 2006). These costs of mobile video telephony can be reduced by technical innovations, and reducing the costs can be an important solution to increasing the use of the extra visual modality in communication. However, reducing the costs does not seem to have done the trick for mobile video telephony, indicating the current benefits might not be as salient as expected. Therefore this paper will elaborate on the possible benefits of an extra visual modality, considered within the constraints that stem from the current mobile telephones.

Research on multimodal perception gives an overview of the benefits of the visual modality for communication. Many studies have shown an advantage of multimodal perception over unimodal perception. For example, the visual-auditory modalities are complementary on individual phonetic features (Robert-Ribes, Schwartz, Lallouache, & Escudier, 1998) and are synergetic: combined auditory-visual perception is superior to perception through either audio or vision alone (Erber, 1975); i.e., the whole is more than the sum of its parts. This synergy is especially salient in noisy

surroundings, where the bimodal advantage can become as large as a 39% increase on intelligibility (Risberg & Lubker, 1978). Effects on memory and emotion have also been shown: adding visual gestures to auditory speech improves the quality of the memory for speech (Kelly, Barr, Church, & Lynch, 1999). Moreover, both modalities supply complementary information about emotions, which are effectively combined (De Gelder & Vroomen, 2000). The latter can be seen as an explanation for the main use of mobile-video telephony for social and emotional talk. A larger body of evidence thus supports the notion that multimodal communication is superior to unimodal communication.

Seen within the context of the mobile device, some restrictions may be imposed on these bimodal advantages. This might prevent mobile video telephony from obtaining the wanted optimal levels of communication as with the reported high levels of synergy for intelligibility, or the mentioned advantages for emotional communication (De Gelder & Vroomen, 2000) and task efficiency (Kraut et al., 2002). Several factors are expected to be of influence, among which are bimodal issues (i.e., spatial coherence, source coherence and temporal coherence), auditory issues, and visual issues. This triplet of factors is discussed in the following paragraph.

### **1.1 Issues in bimodal communication**

The next section will give an overview of the issues which might prevent mobile video telephony from obtaining optimal levels of communication concerning intelligibility, emotion or task efficiency, starting with the bimodal issues of spatial coherence, source coherence and temporal coherence.

Spatial coherence; i.e., a heard sound is attributed to a seen object. This is also known as the ventriloquism effect (Vroomen & De Gelder, 2004) and should be sufficient in order to ascribe the heard speech to the seen face. In the most common usage setting, with the image at the same location as the sound, this will almost always be the case. In less common settings, for example when the user is wearing

a single-ear headset, this effect may disappear, though the effect is usually found to be very robust (Bertelson, Vroomen, & De Gelder, 1997).

Source coherence; i.e., two sources behave in a way that a heard stimuli is ascribed to them both. This coherence should be sufficient in order to prevent a McGurk effect (McGurk & MacDonald, 1976) from occurring. This effect occurs when the visual source and the auditory source do not behave equally (e.g., the lips move different from the sound). This is unlikely to become a problem, since the modalities are both taken from the same source. Though exceptions can exist, for example when the image quality is reduced due to connectivity loss.

Temporal coherence; i.e., different events take place at (almost) the same time and are thus seen as one stimuli, between the auditory and visual modalities should be within sufficient ranges. Summerfield (1992) concludes an auditory delay of at most 80ms has no effects on intelligibility, though researchers differ on the amount of auditory delay that has an effect (e.g., Dixon & Spitz, 1980). Most differences come from different stimuli and methods. These differences highlight the possibility of a specific temporal range for mobile video telephony, for example due to an interaction between screen size and the just-noticeable-difference level of video delay. Because of technical difficulties a delay may occur quite often for either modality, which makes the temporal range very relevant for achieving optimal bimodal communication. Depending on the coding methods, traffic congestion, and network capacity (i.e., speed) the video delay is general between a few tens to several hundreds of millisecond (Basso, 2006).

Within the auditory issues, auditory stimulus quality is often a problem which prevents the mobile video telephone from being used in noisy surroundings like bars (O'Hara et al., 2006), while in these settings the bimodal synergy should be most salient. Many specific factors influence the auditory stimulus quality, such as the range of frequency or the volume.

A comprehensive review of these factors is beyond the scope of this paper.

Concerning the issues specific to the visual modality, the following are expected: temporal frequency, spatial resolution, screen size, noise and zoom-level. Temporal frequency and spatial resolution can be at low levels before decreasing intelligibility: a 25Hz temporal frequency (Vitkovitch & Barber, 1996) and 32x32 spatial resolution (Brooke & Templeton, 1990). But these experiments leave out the relationship with the screen size, which influences the total image quality. A small spatial resolution might be, dependable on the screen size, better explained as a noisy stimulus. Noise has also been investigated by blurring the visual stimulus of a talking face; several experiments showed visual speech perception was not impaired until a certain blurring threshold was reached (Thomas & Jordan, 2002). Zoom-level determines the parts of the face that are visible, and is related to intelligibility: the more of the face is visible, the greater the intelligibility (Benoît, Guiard-Marigny, Le Goff, & Adjoudani, 1996).

Of all these visual issues, most of the effects have already been investigated. Not mentioned in this overview are the effects of screen size. In cooperation with this study, these effects have been investigated as well. Screen size is a very salient issue since a small screen size is a key feature of mobile phones. Screen size was found to have a significant effect on the intelligibility of bimodal messages. Furthermore, in determining the boundaries of optimal bimodal communication, screen size is expected to produce interaction effects on other factors as well, due to its key relevance in bimodal communication.

Defining screen size is not as straightforward as one might expect. Important characteristics of screens that need to be incorporated are Field Of View (FOV), Pixel Per Inch (PPI), resolution and the actual physical size (display size). When discussing the effects of screen size, these characteristics might each have a separate influence on the measured effect. FOV is the size of a screen relative to the eye

of the user, thus 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. FOV has been found to enhance the spatial abilities of the users, till certain optimal levels (Tan, Gergle, Scupelli, & Pausch, 2006). 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 (Brooke & Templeton, 1990). Resolution is related to PPI, as PPI can be computed from the display size and its resolution. Finally, the physical size of screens is expected to influence the effects as well, through a “bigger is better” rule (Josephs, Giesler, & Silvera, 1994).

As said above, besides the characteristics of screen size, numerous interactions between other bimodal issues and screen size may exist. For example, smaller screens might allow for a larger temporal latency. Since temporal latency relatively frequently occurs in mobile video telephony, such an interaction is of specific interest within this context. But, since there is an infinite amount of possible interactions, it is infeasible to examine all of them. Though, based on the effects found in the cooperating study of the effects of screen size on intelligibility, several possible interactions can be predicted. Among which, the relation between screen size and temporal latency.

## 1.2 Hypothesis and predictions

One of the main reasons for using mobile video telephony should be in the benefits that it gives to the user. These benefits can particularly be found in the synergy created by using both the auditory and the visual modalities. This synergy can become very large, but does not seem to be achieved by mobile video telephony.

The visual modality becomes less important with decreasing screen size, as shown in the cooperating study for intelligibility. Considering the context of mobile video telephony and the importance of a possible interaction effect with temporal latency, it can

be expected that temporal latency becomes less of a problem with a decreasing importance of the visual modality. This results in the following hypothesis;

R1: An interaction between temporal latency and screen size exists, such that a larger temporal latency may exist with a smaller screen size, before decreasing intelligibility.

In order to answer this hypothesis, the importance of the visual modality relative to the auditory modality should be increased. This is generally done by decreasing the reliability of the auditory channel by adding noise to it. By doing so, any changes in the visual channel will produce a greater effect on the final intelligibility. For this, the optimal level of to-be-added noise will be a secondary research question;

R1.1: Which signal-to-noise ratio(SNR) for the auditory modality maximally enhances the importance of the visual modality?

In the following chapters these hypothesis will be tested using two separate studies; using a pilot study to identify the optimal SNR (R1.1) and an experimental study (Main Study) to test the main research hypothesis (R1). The methods and results for these studies will be described and discussed.

## **2. Method**

The following section describes the method by which the hypotheses were tested. First the method of the pilot study will be discussed, after which the main study will follow.

### **2.1 Pilot study**

A pilot study was conducted to answer R1.1: Which SNR for the auditory modality maximally enhances the importance of the visual modality?

In this section, the pilot study will be described. The SNR is thought to be optimal when the largest synergy effects are visible. The pilot study consists of two stages that are described in this method section. After a test on intelligibility, the size of the synergy of the visual information can be estimated.

#### *2.1.1 Participants.*

In the pilot study six people participated. All participants were students of the University of Twente with normal or corrected to normal vision and hearing. All participants participated on a voluntary basis. The participants of the pilot study were not allowed to participate in one of the main studies. This precaution was necessary to prevent learning effects from the experience with the pilot study from influencing the results of the main studies.

#### *2.1.2 Materials and Apparatus.*

A computer with accompanying headset was used to present the videos with audio or the audio alone. All videos were shown on a 15" computer screen. The participants who were participating in the audio-alone condition of the second part of the pilot study saw a white screen during the experiment instead of the videos. In the first part of the pilot study, all participants saw a white computer screen when they were listening to the audio.

The listening test that was used in the first part of the pilot study is the first half of an official English listening test normally used on university preparatory secondary education in the Netherlands to indicate the knowledge level of the English language. The test contains standardized questionnaires and norm based scoring forms. The original norm based correction materials were used to determine the level of knowledge of the English language of the participants.

A set of 9 videos was selected from the original material. The videos that were selected had the most footage where the face of the person who was talking in the videos was visible on screen. This was done because the experimenters thought that 'disturbing' screen shots, in which the face of the talking person was not visible, could negatively influence the intelligibility of the audio signal.

For the second stage of this pilot study, different SNR values were added to the material. According to different literature the used SNR usually differs between -0dB and -24dB, with a maximum synergy around -18dB (Benoit, Mohamadi, & Kandel, 1994).

- (1)  $\text{RMS}_{\text{amplitude,signal}} = 20 * \log_{10} \frac{X_{\text{signal}}}{X_{\text{ref}}}$
- (2)  $\text{RMS}_{\text{amplitude,noise}} = 20 * \log_{10} \frac{X_{\text{noise}}}{X_{\text{ref}}}$
- (3)  $\text{SNR}_{\text{dB}} = 20 * \log_{10} \frac{X_{\text{signal}}}{X_{\text{noise}}}$
- (4)  $\text{SNR}_{\text{dB}} = (\text{RMS}_{\text{amplitude,signal}} - \text{RMS}_{\text{amplitude,noise}})$

Figure 1. Computation of the SNR.

However these effects are found with a stimulus of 18 words. Stories are harder to fully understand and probably have other synergy levels at other ratios. For this reason, the ratios tested in the pilot study are lower than in the described literature, namely; -12dB, -9dB and -6dB.

The SNRs are defined as described in Figure 1. The Root Mean Square (RMS) amplitudes of the different sources (signal and noise) were computed in decibel Full Scale (dBFS). This dBFS is always negative, and is defined as a logarithmic scale with as reference the clipping point of the hardware. Thus, 0dBFS is the loudest possible output. The RMS value is a logarithmic function of the power of the source ( $X_{\text{signal}}$  and  $X_{\text{noise}}$ ) and the power of the reference point (equation 1 and 2). The SNR (in dB) then is defined as a logarithmic function of the power of the different sources (equation 3). These three functions can be rewritten to equation 4.

The RMS values of the signals are changed in order to change the SNR. These values were -12dB, -15dB and -18dB. The RMS value of the noise was kept constant at -6dB. The three possible combinations resolved in SNR values of -6dB, -9dB and -12dB. These SNRs were used in this pilot study.

The type of noise used among various studies differs. Most studies use white noise (e.g., Benoit et al., 1994; Robert-Ribes et al., 1998); an equal amount of energy that is added to all frequencies. Other studies use more sophisticated types of noise (e.g. Drullman, 1995), such as pink noise. With this latter type of noise a decreasing amount of energy is added in the higher frequencies. In this

experiment white noise was used because this type was thought to be the most effective one and the one that is most realistic to occur in mobile-video phone use.

### 2.1.3 Procedure

As said above, the pilot study consisted of two stages. The first stage tested the participants on their knowledge of the English language, with an audio test on a computer in normal screen size. With this audio test the original questionnaire was used to get an accurate result. The described questionnaire measured how well the participants understood the content of the presented material. The same audio and accompanying test was presented to all participants. This was done to make it possible to find out if the amount of knowledge of the English language influences the score on the second part of the main study.

After this first part of the pilot study, pairs were made and each of these 3 pairs got test material in the second part of the pilot study with another SNR.

After the first stage, every participant was invited for the second stage. The second stage used also a normal screen size and no forced distance to the screen. For this stage, each pair was presented with stimuli with one of the 3 SNR values. One halve of the pair saw and hear the material, the other only heard the material. Each halve of the pair will do the test separately from the other halve. In sum, the second stage of the pilot study will consist of six conditions: 3 (SNR) X 2 (with or without video).

Before the experiment demographic data was asked like name, sex, age, occupation and nationality. Before each test participants was told that he or she is going to watch a video or hear a story which is an English listening test, designed to measure their understanding of the stimulus. Participants were asked to perform as well as possible on the test. Participants were asked if they need to and were wearing any glasses, lenses or hearing-aid. For the second stage of this pilot study, it was told that noise has been added to the auditory material.

## 2.2 Main study

In this section will be described how the main study was constructed and carried out.

With this experiment the researchers wanted to examine the second hypothesis of this paper (R2): An interaction between temporal latency and screen size exists, such that a larger temporal latency may exist with a smaller screen size, before decreasing intelligibility. First something will be told about the participants of the experiment. The other two sections are about the materials that were used for the experiment and finally the whole procedure of this first main study is pointed out.

### 2.2.1 Participants

Every participant watched the three videos in a certain order on the three different screen sizes. This gave 36 possible unique conditions. Before the experiment these conditions were constructed and 11 of them were used. Almost all participants were students of the University of Twente. Participants participated on a voluntary basis or received study credits for their participation in the experiment. The latter happened if the participant was a first or second grade Psychology student, because only students of these cohorts can collect these credits. In this experiment 8 participants received study credits.

### 2.2.2 Materials and Apparatus

Because of the results of the pilot study and findings in literature was decided that a SNR of -9dB was used in the main studies. This SNR was computed as described in the method section of the Pilot Study.

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 and is described here.

Because the distance participants had to the screen could influence the results of the experiment, they were forced to keep a particular distance and position by placing the chair in front of the computer and the computer keyboard on a fixed location. As

mentioned in the Introduction during a conversation using a mobile video telephone, latency can occur. In this second study a latency of 250ms was added to the audio of the official CITO English video/listening task; i.e., the lip movements preceded the sound 250ms. Furthermore a transparent screen was placed in front of the computer monitor that prevents the participants from coming too close to the monitor.

There was also the possibility to watch participants during the experiment with a camera to see if participants kept the right distance and position. This was used to check if people were not moving their body too much during the experiment.

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 screen relatively larger for the eye and thus increases the FOV, which might diminish any effects of difference in intelligibility between the three screen sizes.

As mentioned in the Introduction during a conversation using a mobile video telephone, latency can occur. In this study a latency of 250ms was added to the audio of the official CITO English video/listening task; i.e., the lip movements preceded the sound 250ms.

A computer was used to show a video with delayed audio. Participants heard the noisy sound of the video message through professional headphones. Three screen types were used, namely; computer screen size, PDA screen size and mobile (video) phone screen size. Table 1 together with the size of the screens, resolution of each screen size and the FOV. In literature, different definitions of FOV can be found. In this research the FOV is the angle subtended from the eyes to the left and right edges of the displayed screen (Tan, 2004).

All these screen sizes were shown on the same computer monitor. To prevent differences in screen quality to influence the results, this was kept as stable as possible

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
Mobiel (video) phone	6.02	177x158	1.12

throughout the different screen sizes by maintaining a screen resolution of 1280x1960 during the experiments.

### 2.2.3 Procedure

The experiment consisted of 36 different conditions. To show the videos of the English video-audio test, three different screen sizes were used. Before the experiment, participants were told that they are going to watch a video which can be seen as part of a video/listening test. Participants were asked to pretend as if it was a real video/listening test and to remember as much as possible from the video.

In each condition, participants saw the three videos on a computer screen in one of the three sizes. For each of the first 36 participants, the combination and/or order of a certain video on one of the screen sizes will be different. So, before the experimentation starts the 36 unique test conditions were programmed. The temporal latency of 250ms was added to these conditions. Participants were randomly assigned to one of the conditions. After each video they had time to fill out the question that belongs to that video. The duration of the experiment was approximately 20 minutes but depended also on how long it took for a participant to answer the questions.

All the videos contain sound with the -9dB SNR ratio that was investigated in the pilot study . As mentioned before a latency of 250ms was added to the audio of the official CITO English video/listening task. At the beginning of the experiment some demographic data was asked; i.e., name, sex, age, occupation and nationality. Although participants need to have normal vision or hearing or corrected to normal vision or hearing, they also had to fill out if they were

wearing their glasses, lenses or hearing-aid. The experimenters told everyone who wanted to participate that it was obligatory to wear the mentioned hearing or vision aids if necessary.

The speech in the video was played with decreased auditory quality trough professional headphones. The volume of the audio was fixed. As mentioned before a transparent screen in front of the computer monitor, and a fixed chair and computer keyboard prevented the participant of coming too close to the computer screen and moving too much.

After the videos the participants had to answer some questions about the difficulty of it. These questions were asked to evaluate how the participants experienced the experiment. According to research literature (Osgood, Suci, & Tannenbaum, 1957), the subjective experience with media can be divided in three factors that influence the evaluation of this experience; evaluation, activity and potential. The evaluation factor evaluates if an experience was pleasant/unpleasant or interesting/not interesting. The factor of activity can be measured with questions about the way in which participants had to be 'active' or felt excited or not during the experiment. The last factor; potential, aims to measure to what extent the participant experienced the experiment as tough or difficult. These three factors underlie the questions that participants had to answer at the end of the experiment to give their evaluation of the experiment. The questions were in a multiple choice format.

Participants were asked to tell nothing about the experiment to other students. The test results, rated with the original CITO scoring forms, were used to determine the rate of intelligibility of the videos. During the experiments answering time was measured over and across different screen sizes.



### 3. Results

In the following, the results will be discussed for each study.

#### 3.1 Pilot study

In the pilot study two tests were carried out; the first test aimed to measure the English language knowledge of the participants and the second test measured the intelligibility of the official CITO video listening test, with or without video. Table 2 shows the results of this Pilot Study, where the second and fourth column contain the number of correct answers on the first and second test.

In the third and fifth column the percentage of correct answers on the first and second test can be found. The different conditions in the experiment are here represented by a number ranging from 1 to 6 that can be found in the first column. In all conditions the first test was an English listening task, the second one an audio only or audio-video test with one of the three chosen signal-to noise ratios. For each condition is mentioned if it was an audio only- or an audio and video condition and what the height of the SNR was.

Only from the conditions with a SNR of -6dB can be concluded that the video increases the intelligibility of the test material, although it is also possible that this effect is partly caused by the better knowledge of the English language of the participant in the second, video-audio condition. This is one of the reasons why these SNR was not chosen for the main study. Another reason is that in comparison with the 3-6 conditions the score on the 2<sup>nd</sup> test in the first and second condition was rather high especially in the video-audio condition and therefore the experimenters thought that this SNR could maybe be too high to show a significant effect of synergy in the main studies.

In the third and fourth condition, there was a very big difference between the scores on the English listening task. As visualized by Table 2, the participant in condition 3 had only one out of 9 correct answers more than the one in condition 4. This difference is very small in

Table 2.

*Summary of the descriptive statistics of the Pilot Study*

Condition	SNR (dB)	Accuracy (first test)		Accuracy (second test)	
		Score	Percentage	Score	Percentage
1. audio	-6	8/12	66.67	4/9	44.44
2. video	-6	10/12	83.33	7/9	77.78
3. audio	-9	12/12	100.00	4/9	44.44
4. video	-9	3/12	25.00	3/9	33.33
5. audio	-12	7/12	58.33	5/9	55.56
6. video	-12	9/12	75.00	5/9	55.56
Mean			68.06		51.85

comparison with the difference between the scores of the English knowledge task and this can possibly show that the synergy of the video is high. Therefore and also because of the results from literature the experimenters choose to use the SNR of these two conditions for the main studies.

The reason why the SNR of -12dB was not chosen for the main studies is that there was no difference found in scores on the third test between the fifth and sixth condition.

#### 3.2 Main study

The main study was designed to investigate the effects of screen size on intelligibility of a bimodal message with an audio delay of 250ms.

The reliability of the measurement devices has not been analyzed, since the devices have already been tested and evaluated in previous studies using large pools of participants as a norm group (N=359)<sup>1</sup>. Considering the within-subjects design, the number of participants for each condition was too small to perform a statistical analysis.

<sup>1</sup> Tested by CITO. See [www.cito.nl](http://www.cito.nl) for more information about this organization.

Table 3. Adopted from Van der Sluis et al. (2007).

*Accuracy scores on each question for two screen sizes and the difference between the small and large screen sizes*

Question	Accuracy			
	Mean	Screen Size		Large-Small
		Small	Large	
1.1	0.70	0.67	0.61	-0.06
1.2	0.48	0.44	0.56	0.11
1.3	0.57	0.44	0.72	0.28
2.1	0.33	0.28	0.33	0.06
2.2	0.83	0.72	1.00	0.28
2.3	0.70	0.72	0.67	-0.06
3.1	0.76	0.61	0.76	0.15
3.2	0.81	0.89	0.67	-0.22
3.3	0.93	0.94	1.00	0.06
Mean	0.68	0.63	0.70	0.07

Although the best 9 videos were chosen for this study, the experimenters thought that three of these 9 videos could possibly influence the outcomes negatively because of irrelevant or false footage. The mentioned footage could confuse the participants and influence their score negatively, because the footage is in place of the speaker at critical moments or is relevant to the storyline but false. False footage has been found to be able to have a negative influence on the outcomes of listening assignments such as these (Gruba, 1997). For these reasons the validity of each specific question has been checked performing a descriptive analysis on the difference in accuracy between the smallest and largest screen size. As can be seen in Table 3. The source for analyzing the validity of the questions was a cooperating study which used 54 participants. Thus; the in Table 3 named accuracy scores are based on another study which used the same material as this one.

The mentioned results of the accompanying study are also used to compare the results with the in the accompanying- and this study used material and the statistics of the norm group results. These statistics allow a comparison with the mentioned norm group results, supporting the previous results reliability of the pilot study and thus research question 1.1, concerning the effects of the SNR of -9dB.

This comparison has been made using a t-test on the means of both results concerning all 9 questions. This t-test showed a significant difference between the norm group results (M=7.58, SD=0.87) and the current results (M=6.13, SD=1.33),  $t(53)=-7.99$  ( $p<.001$ ).

The questions that had an uncertain validity performed different (i.e., had a negative difference) compared to the other questions (i.e., which had a positive difference). Because the data underlines the possible theoretical problems raised about the validity of these questions, they were left out of any further analysis in the mentioned cooperating study and also in this main study.

The primary effect of screen size on intelligibility was analyzed using an analysis of variance (ANOVA) on accuracy and answering time by screen size for all videos. The accuracy for each video, used as an indicator of intelligibility, was computed by summing the accuracy of the questions specific to that video. The same was done for answering time. The descriptive statistics of this summation are shown in Table 4 and Figure 3. All statistical tests are performed with an alpha level of .05.

The analysis of variance on accuracy was found to be non-significant, indicating no influence of screen size on intelligibility in the

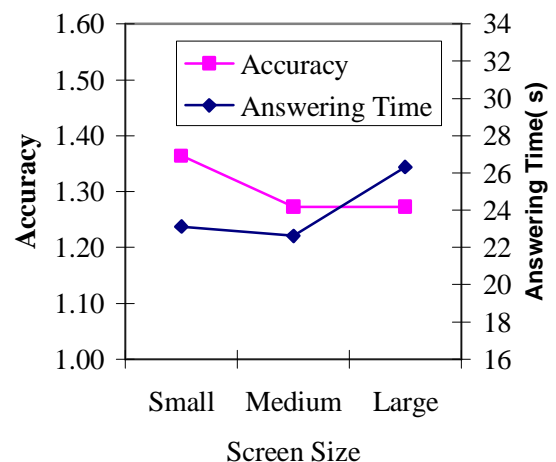


Figure 3. Accuracy and answering time for the different screen sizes.

Table 4.

*Descriptive statistics for accuracy and answering time for each screen size, video and sequence*

Screen Size	Accuracy		Answering Time	
	Mean	Std. Deviation	Mean	Std. Deviation
Small	1.36	0.67	23.12	3.46
Medium	1.27	0.47	22.61	12.36
Large	1.27	0.65	26.32	4.07
Video				
1	1.09	0.70	27.20	9.94
2	1.45	0.52	20.70	10.71
3	1.36	0.50	24.15	15.45
Sequence				
First	1.18	0.60	27.03	15.24
Second	1.18	0.60	27.13	10.60
Third	1.55	0.52	17.89	8.30
Mean	1.30	0.59	24.02	12.19

expected direction. For answering time the same was found. The other expected effects are sequence and video effects and difference in difficulty per video. The analysis was done using an ANOVA on accuracy and answering time. The descriptive statistics of these analyses are also shown in Table 4.

For sequence the analysis of variance on accuracy and answering time was found to be non-significant. For video the results were also found to be non-significant although the found results for variance in answering time on video suggested a trend.

#### 4. Discussion

The discussion consists of two parts. First will be a review of the study and its results, accompanied with a thorough discussion about it. The second part will discuss the implications of this study.

##### 4.1 Interpretation of the findings

This study explored the interaction of screen size with temporal latency between the auditory and visual stimuli. For this study, the auditory stimulus quality has been reduced by decreasing the SNR to -9dB, based on findings from the Pilot Study and other studies. The results of the Pilot Study indicated a large increase in intelligibility starting from a SNR of

-6dB. This does not fully answer research question R1.1, asking which SNR maximally enhances the intelligibility. But, together with findings from others studies, does contain enough information to base a sufficient SNR on, namely -9dB. Furthermore, the results on the video-listening test of the Main Study confirmed the influence of the SNR on the performance on the test with a significant decrease in accuracy

scores of 16.11% compared to the norm scores of the test.

The main study explored the main hypothesis (R1). This was tested by presenting the auditory stimulus with the SNR of -9dB and the visual stimulus at three different screen sizes (mobile phone, PDA and monitor), and measuring intelligibility of these bimodal messages by using a standardized video-listening test. Furthermore, a temporal latency of 250ms was added between the visual and the auditory stimuli. Due to relatively few participants (N=11), none of the findings were significant. However, the data did indicate an effect in the opposite direction as found in the cooperating study. To be precise, a larger screen size produced lower accuracy scores.

These results suggest that the interaction effect predicted by the main hypothesis (R1) is correct. The problems of temporal latency are enhanced by a larger screen size, indicating a more intense processing of the stimuli from the larger screen. The findings on answering time support this as well, being that the largest screen size produced a higher answering time, thus required more processing.

These results are very interesting and relevant for mobile (video) phone use. The used

temporal latency, 250ms, has been found to cause an effect in other studies as well (Summerfield, 1992), and to occur quite often in common everyday usage (Basso, 2006). Although a minor effect was found it is imaginable that another temporal latency would have given other results. Therefore in future studies the effect of different amounts of temporal latency on different screen sizes should be investigated.

#### **4.2 Implications of the findings**

The implications of the findings from the studies will be discussed two-folded; at a theoretical level and at a practical level.

The results of the cooperating studies suggested a difference between two processing stages. This notion is supported by several cognitive models as well. For example, Wickens (2002) identifies four different stages of processing. The first two are perceptual and cognitive activities, which can be translated to the integration and processing of the stimuli such as identified by the results. Furthermore, the ACT-R cognitive architecture (Anderson et al., 2004) also differentiates between a visual module for perceiving the environment, and a pattern matcher for interpreting the perceived stimuli. Thus, different stages of processing information exist: one at a perceptual level, and one at a cognitive processing level.

The indications of the main study, on the interaction between screen size and temporal latency, suggests some interaction between both stages. Since in this study the intelligibility decreased with a larger screen size, it can be concluded that a failing bimodal integration (due to temporal latency) is more problematic to the cognitive processing of the stimuli (indicated by the intelligibility), especially when the modalities contribute more equally to the perception (in cases of a larger screen size).

Also, on a theoretical level, the lacking difference between men and women is striking with most other findings from media psychological experiments concerning the influence of screen size (e.g., Grabe, Lombard, Reich, Bracken, & Ditton, 1999). The main difference between the studies from most

media psychological scholars and this one is the used screen size. This experiment focused on relatively small screens, where the gender differences are mostly found for large screens. Furthermore, the differences are mainly at other effects (e.g., arousal). This indicates that the difference between smaller screens works through other mechanisms than the difference between larger screens. This is confirmed by findings that effects on arousal by large screens interacts with media content and gender (Grabe et al., 1999), suggesting these effects occur at a cognitive level instead of a perceptual level.

Beside of the theoretical implications, there are several practical issues that have to be taken into account. Knowing that screen size influences the communication through mobile video telephony at a very basic level asks for future research about the boundaries of this effect and about higher level effects. Furthermore, it asks for ways to cope with these effects.

This holds for the temporal latency effects which were explored by the Main Study. It seems the strength of temporal latency effects interacts with other factors such as screen size. Further research has to examine under which conditions which temporal latency is disturbing. As mentioned in the Introduction, in literature there is no consensus about the smallest amount of temporal latency that is noticeable and disturbing for people. Therefore it is impossible, at least at this point, to give strict guidelines about the amount of temporal latency between the audio and video that is acceptable for mobile video telephony in general.

Another practical application that could be derived from the in this paper described research is that there probably is a threshold for which video still contributes to the intelligibility of the audio. This can be very useful information for users since when the video does not have a surplus value it is more practical to turn it off and instead use only audio. To investigate how this could be implemented in mobile-video phones and under

which conditions the screen does not have a surplus value anymore need to be investigated in further research.

Knowing such a threshold for temporal latency can give rise to technical solutions which increase the chance the latency stays below this threshold. Using a buffer set some time before the threshold, the occurrence of increased latency can be captured with the buffer. By this, the occurrence of temporal latency issues can be decreased without notice of the user, by slightly increasing the overall (unnoticeable) temporal latency.

In the above text are different practical restrictions, possibilities of and solutions for mobile-video phone use addressed. Most of these were not or only partly investigated by the experiments of this paper. Therefore the experimenters see it as a challenge for future research to investigate the restrictions and possibilities of mobile-video phone use in relation to different multimodal effects as they are described in the introduction of this paper.

## 5. References

- Anderson, J. R., Bothell, D., Byrne, M. D., Douglass, S., Lebiere, C., & Qin, Y. (2004). An Integrated Theory of the Mind. *Psychological Review*, *111*(4), 1036-1060.
- Basso, A. (2006). Beyond 3G video mobile video telephony: The role of 3G-324M in mobile video services. *Multimedia Tools and Applications*, *28*(1), 173-185.
- Benoît, C., Guiard-Marigny, T., Le Goff, B., & Adjoudani, A. (1996). Which components of the face do humans and machines best speechread? In D. G. Stork & M. E. Hennecke (Eds.), *Speechreading by humans and machines: Models, systems, and applications* (pp. 315-328). Berlin: Springer-Verlag.
- Benoit, C., Mohamadi, T., & Kandel, S. (1994). Effects of Phonetic Context on Audio-Visual Intelligibility of French (Vol. 37, pp. 1195-1203).
- Bertelson, P., Vroomen, J., & De Gelder, B. (1997). *Auditory Visual Interaction in Voice Localization and in Bimodal Speech Recognition: The Effects of Desynchronization*. Paper presented at the Proceedings of the Workshop on Audio-visual Speech Processing: Cognitive and Computational Approaches, Rhodes, Greece.
- Brooke, N. M., & Templeton, P. D. (1990). Visual speech intelligibility of digitally processed facial images. *Proceedings of the Institute of Acoustics*, *12*, 483-490.
- De Gelder, B., & Vroomen, J. (2000). The perception of emotions by ear and by eye. *Cognition and Emotion*, *14*, 289-311.
- Dixon, N. F., & Spitz, L. (1980). The detection of audio-visual desynchrony. *Perception*, *9*, 719-721.
- 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.
- Erber, N. P. (1975). Auditory-visual perception of speech. *Journal of Speech and Hearing Disorders*(40), 481-492.
- Grabe, M. E., Lombard, M., Reich, R. D., Bracken, C. C., & Ditton, T. B. (1999). The role of screen size in viewer experiences of media content. *Visual Communication Quarterly*, *6*, 4-9.
- Gruba, P. (1997). The role of video media in listening assessment. *System*, *25*(3), 335-345.
- Josephs, R., Giesler, R., & Silvera, D. (1994). Judgment by quantity. *Journal of Experimental Psychology: General*, *123*(1), 21-32.
- Kelly, S. D., Barr, D. J., Church, R. B., & Lynch, K. (1999). Offering a Hand to Pragmatic Understanding: The Role of Speech and Gesture in Comprehension and Memory. *Journal of Memory and Language*, *40*, 577-592.

- Kraut, R. E., Gergle, D., & Fussell, S. R. (2002). The use of visual information in shared visual spaces: informing the development of virtual co-presence, *Proceedings of the 2002 ACM conference on Computer supported cooperative work*. New Orleans, Louisiana, USA: ACM Press.
- McGurk, H., & MacDonald, J. (1976). Hearing lips and seeing voices. *Nature*, 264(5588), 746-748.
- O'Hara, K., Black, A., & Lipson, M. (2006). Everyday practices with mobile video telephony. *Proceedings of the SIGCHI conference on Human Factors in computing systems*.
- Osgood, C. E., Suci, G. J., & Tannenbaum, P. H. (1957). *The Measurement of Meaning*. Urbana: University of Illinois Press.
- Risberg, A., & Lubker, J. (1978). Prosody and speechreading. *Speech Transmission Laboratory Quarterly Progress Report and Status Report*(4), 1-16.
- Robert-Ribes, J., Schwartz, J. L., Lallouache, T., & Escudier, P. (1998). 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), 3677-3689.
- Short, J., Williams, E., & Christie, B. (1976). *The social psychology of telecommunications*. London: Wiley.
- Summerfield, Q. (1992). Lipreading and Audio-Visual Speech Perception. *Philosophical Transactions: Biological Sciences*, 335(1273), 71-78.
- Tan, D. S. (2004). *Exploiting the cognitive and social benefits of physically large displays*. Carnegie Mellon University, Pittsburgh, Pennsylvania.
- Tan, D. S., Gergle, D., Scupelli, P., & Pausch, R. (2006). Physically large displays improve performance on spatial tasks. *ACM Transactions on Computer-Human Interaction*, 13(1), 71-99.
- Thomas, S. M., & Jordan, T. R. (2002). Determining the Influence of Gaussian Blurring on Inversion Effects With Talking Faces. *Perception & Psychophysics*, 64, 932-944.
- Van der Sluis, F., Van Drunen, A., & Van den Broek, E. L. (2007). The effects of screen-size on intelligibility in multimodal communication. Enschede: Twente University.
- Vitkovitch, M., & Barber, P. (1996). Visible speech as a function of image quality: effects of display parameters on lipreading ability. *Applied cognitive psychology*, 10(2), 121.
- Vroomen, J., & De Gelder, B. (2004). Perceptual effects of cross-modal stimulation: The cases of ventriloquism and the freezing phenomenon. In G. Calvert, C. Spence & B. E. Stein (Eds.), *Handbook of multisensory processes* (pp. 141-150). Cambridge: MIT Press.
- Whittaker, S. (2003). Things to Talk About When Talking About Things. *Human-computer interaction*, 18(1 & 2), 149.
- Wickens, C. D. (2002). Multiple resources and performance prediction. *Theoretical Issues in Ergonomics Science*, 3, 159-177.