KTH Tal, musik och hörsel

EXPERIMENT WITH ASYNCHRONY IN MULTIMODAL SPEECH COMMUNICATION

by

Marie Molander

Centrum för talteknologi

Handledare: Björn Granström, Jonas Beskow
Godkänt den: ____________ Examinator: ____________

Stockholm
Juni 2003

Examensarbete i Talteknologi

Institutionen för tal, musik och hörsel
Kungliga Tekniska Högskolan, 100 44 Stockholm

KTH Tal, musik och hörsel

Examensarbete i Talteknologi

Experiment med asynkroni i multimodal talkommunikation

Marie Molander

Godkänt År, mån, dag	Examinator	Handledare

Sammanfattning

Målet med den här studien var att undersöka tidsfördröjningseffekter vid audiovisuell talperception för naturliga och syntetiska ansikten. Huvudfokus var på SYNFACE-projektet, utvecklingen av ett hjälpmedel för hörselskadade vid telefon-kommunikation.

Det som undersöktes i experimenten var hur vi påverkas av tidsförskjutningar mellan ljud och bild. Ljudkanalen bestod av naturligt tal med en vocoder-liknande störning för att simulera hörselskada. Tolv olika försöksvillkor presenterades för försöks-personen vid två separata tillfällen. Det naturliga ansiktet testades för försöksvillkor där ljudet kom före bilden (negativa siffror), såväl som efter (positiva siffror), medan det syntetiska ansiktet endast testades för det förra fallet. De asynkronier som undersöktes var 50, 175 och 300 ms. Dessutom undersöktes två referensvillkor: synkroni och audio-only (dvs endast ljud).

ANOVA-tester för båda ansiktena avslöjade att varken -300 ms eller -175 ms var signifikant bättre än audio-only, vilket indikerar att den slutgiltiga SYNFACE-produkten inte skulle vara till hjälp när det gäller så stora tidsfördröjningar. Däremot visade -50 ms ingen signifikant försämring jämfört med synkroni. Tyvärr är fördröjningen i den nuvarande SYNFACE-prototypen större än detta. Det skulle därför vara intressant att undersöka asynkronier mellan -175 ms och -50 ms för att veta exakt var uppfattbarheten försämras. ANOVA visade även att ansiktstypen inte har

någon effekt, vilket tyder på att kvaliteten av det syntetiska ansiktet är nära nivån för det naturliga ansiktet.

Toleransen för då ljudet kommer efter bilden är större än då det kommer före, vilket verifieras av en signifikant prestationsförsämring så sent som vid +300 ms (att jämföra med -175 ms). Det påvisades till och med en *ökning* i uppfattbarhet för +50 ms, vid jämförelse med det synkrona fallet. Denna ökning är dock inte signifikant och efter statistisk analys konstaterades att tidsförskjutningar i intervallet [-50, +175] endast har en liten effekt på ett talat meddelande när det gäller det naturliga ansiktet.

KTH Speech, Music and Hearing	*Master Degree Project in Speech Technology* Experiment with asynchrony in multimodal speech communication Marie Molander
Approved Year, month, day	Examiner Supervisor

Abstract

The purpose of this study was to examine the delay effects in audiovisual speech perception for natural and synthetic faces. The main focus was on the SYNFACE project, the development of a telephone communication aid for hearing impaired persons.

In the experiments, the consequence of temporal displacement of the audio in relation to the visual channel was investigated. The audio channel was natural speech with a vocoder-like distortion to simulate hearing loss. Twelve different experimental conditions were presented to the subjects in two separate sessions. The natural face was tested for audio-leading (negative numbers) as well as audio-lagging (positive numbers) stimuli, whereas the synthetic face was tested only for audio-leading stimuli. Asynchronies examined were 50, 175 and 300 ms. In addition, two reference conditions were examined: synchrony and audio-only.

Tests of ANOVA including both faces revealed that neither -300 ms nor -175 ms were significantly better than the audio-only condition, which implies that the final SYNFACE product would not be beneficial for delays of this magnitude. The -50 ms condition, however, did not show significantly lower intelligibility scores than the synchronous condition. Unfortunately, the delay measured in the present SYNFACE prototype is greater than this. It would, therefore, be interesting to investigate asynchronies between -175 ms and -50 ms to see exactly where the intelligibility drops. ANOVA further showed that the effect of the type of face was non-significant, indicating that the quality of the synthetic face is close to that of a natural face.

The tolerance for audio-lagging delays is larger than for the audio-leading delays, which is verified by a significant decrease in performance as late as at +300 ms (the corresponding audio-leading delay is -175 ms). Even a gain in intelligibility was found for the +50 ms condition compared to synchrony. However, this gain is not significant, and statistical analysis showed that delays within the interval [-50, +175] only have a small effect on the spoken message for the natural face.

Preface

This Master of Science thesis was a part in the development of Synface. The project was performed at the Department of Speech, Music and Hearing (TMH) at the Royal Institute of Technology (KTH), Stockholm, Sweden, and was completed in June 2003. Supervisors were Björn Granström and Jonas Beskow.

Contents

1 INTRODUCTION

This Master of Science thesis is a part of SYNFACE, a project that develops a telephone communication aid for hearing impaired persons. An incoming speech signal controls the movements of a synthetic face that appears on a computer display, thus making speech-reading possible and facilitating conversation for the hearing impaired user. At the present stage, a delay between the sound and the image arises and this is something that can confuse the user. The purpose of the present study is to investigate at what point this delay becomes so big that the benefit of the synthetic face disappears.

The experiments in the present study are made in a broader perspective than that of SYNFACE in order to compare the results to previous studies in the field of audiovisual perception. Researchers have formerly investigated the perception of sentences, words or syllables and how the subjects are affected by asynchrony between the two input channels. This has been done for natural faces as the visual stimulus. In this study we move forward by also investigating the delay effect for a synthetic face.

2 BACKGROUND

2.1 HEARING IMPAIRMENT

The definition of *hearing impairment* is a hearing loss of 21 dB to 94 dB, as reported by Arlinger *et al.* (2003). A person with a loss greater than 94 dB is *deaf*. Hearing losses can occur in several ways. Impairments influencing the transmission through the middle ear can be alleviated by hearing aids that amplifying the sound waves or conducting them to the cochlea through the bone of the skull. Aids for dysfunctions of the inner ear have only been developed over the last years. One example is the cochlear implant, a device stimulating the auditory nerve electrically. However, the attained sensation is far from being comparable with normal hearing. Whether using other types of hearing aids or not, most hearing impaired persons can still be helped by the visual information provided by an application like SYNFACE. Those who cannot, either have a too severe impairment, and do not obtain a sufficient gain from a supplementary face, or a too mild impairment and hear well enough without the face. Comparison with a study performed on normal hearing persons can be made: Agelfors *et al.* (1998) concluded that those with an intelligibility level below 40% or above 80% for the audio-only condition[1] were the ones who would not profit much from the synthetic face.

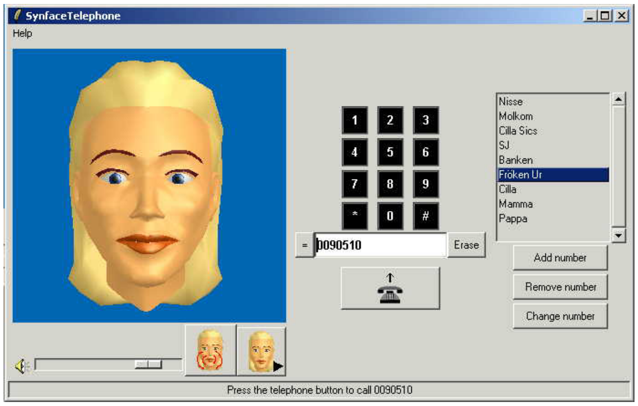
Hearing impaired persons are often experienced speech-readers. In addition to the lips and the tongue they also use other visual information, such as head movements

[1] The audio-only condition often serves as a reference level when investigating multimodal stimuli.

and body language, to achieve a better comprehension of a spoken message. Usually, normal hearing persons also perceive speech in this manner but they might not be aware of it until they find themselves in a noisy environment or exposed to asynchrony on TV. Summerfield suggests that this ability to speech-read could be innate (cited by Schomaker *et al.* 1995). Communication through normal telephones cannot provide this type of supplementary visual information. This can affect the spontaneity and the social benefits typical for face-to-face discussions (Sellen 1992).

2.2 SYNFACE

In October 2001 the former CTT project Teleface was expanded into the EU-project SYNFACE, which is scheduled to proceed over a 3 year period. The full name is ”Synthesized talking face derived from speech for hearing disabled users of voice channels”. A demonstrator interface was developed by Ward (2002) and is shown in Figure 1.



0090510

Nisse
Molkom
Cilla Sics
SJ
Banken
Fröken Ur
Cilla
Mamma
Pappa

Press the telephone button to call 0090510

Figure 1. *The interface of the SYNFACE prototype, developed by Klara Ward.*

For the present SYNFACE prototype the delay is adjustable in order to allow testing of the prototype during the development. This might be possible even for the final product. The choice of this delay has to be a compromise between effectiveness in communication, for which small delays are desired, and optimal speech recognition, which may require a delay of about 100 ms (Pihl 2003). More than 100 ms does not provide a significant improvement in speech recognition quality.

2.2.1 Automatic Speech Recognition

Recognizers are created by using statistical methods (e.g. DP, HMM, ANN) on large amounts of recorded speech in order to allow identification of phonemes. Speech is, however, not made of delimited units but engenders an acoustical flow which evokes reductions and co-articulations of the different phonemes. Pauses, restarts and respiration as well as the variability between speakers also cause problems. Contributing elements to this are accent, sex, age, physiology, but also the present state of health and mood of the speaker. Also, surrounding sounds and noise influence on the system’s level of confidence. The recognizer must be able to handle all these problems and work for improvement is in progress in order to make the system more flexible and less topic-dependent. For SYNFACE, the recognizer's output controls the movements of the talking head through a complex system of parameters responsible for different parts of the face (Beskow 2003).

2.2.2 Talking Heads

Various projects concerning talking heads[2] have evolved throughout the years. These heads can be used not only as an aid for the hearing impaired persons, but also as an interactive language teacher, as informative agents, in games, for multimedia and in dialogue systems.

The animated face for SYNFACE was created by recording a natural face talking, then marking certain reference spots in the face and measuring the displacement of each spot for all articulations. These data serve as the basic information in the development of the talking heads. Several facial models can be chosen in order to adjust the SYNFACE prototype to the user's wishes (see Figure 2).

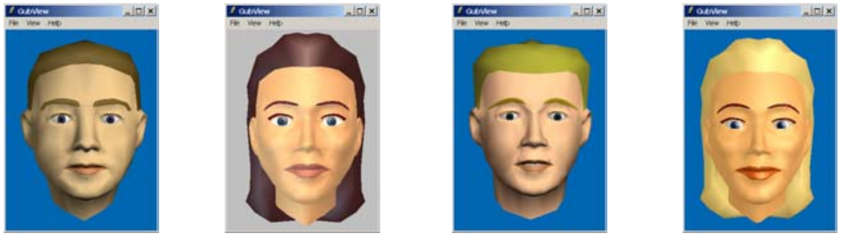


Figure 2. *Examples of facial models used at CTT. From left to right: Alf, Kattis, Sven and Vikthoria.*

Recently, a doctoral thesis treating tongue models, making the tongue movements more informative, was finished by Olov Engwall (2002) at CTT. Work is also in progress to obtain talking heads that are more natural than they are today. This is done by adding emotional gestures and wrinkles as well as eyebrow-, eyelid-, eye- and head movements. It is also possible to adjust the articulation strength.

2.2.3 Turn taking

For the SYNFACE application it might be interesting to introduce a certain delay even for the audio, instead of accepting an asynchrony between the channels. This would produce facial movements that match the voice. Yet, this latency could instead induce problems concerning turn taking. When a person makes a statement, he/she normally expects immediate feedback. But in case of an introduced delay of the entire signal, a latent period arises and misunderstanding of whose turn it is to talk might cause mutual silence or double talk, making communication less effective. In face-to-face communication the speaker can pass along the turn by a gaze or a nod, as expressed by Ruhleder & Jordan (2001), but this is not possible when contact is being made through telephone lines. It has been suggested that some kind of marker could be added to the interface of SYNFACE in order to provide this type of

[2] For links, see http://mambo.ucsc.edu/psl/fan.html.

information. One option is to show a lamp or, maybe even better, to produce specific facial gestures while the audio signal is being received. Then the user would know when to wait for an incoming message and when to talk.

John Tang and Ellen Isaacs examined delay effects in video-conference (Tang & Isaacs 1993; Isaacs & Tang 1993), and found that turn taking is disrupted at a latency of 57 ms, which is of a magnitude crucial for the SYNFACE application. Their subjects were so annoyed by this large latency that they preferred a separate transmission channel for the audio, which instead evoked asynchrony between the two modes. So, they would rather receive the audio nearly instantaneously, despite an asynchronous input and a slight degradation of sound quality (the audio channel from the videoconferencing system was replaced by telephone lines and a speakerphone). This might imply that the introduction of latency would not be desirable for SYNFACE. Tang and Isaacs did not, however, deal with aids for hearing impaired persons and did not have to consider conflicting channels in the same manner. The future users of the SYNFACE product are more dependent on matching channels for the understanding of a spoken message. Other researchers have found thresholds near 57 ms, according to Ruhleder & Jordan (2001).

2.2.4 Competing Products and Projects

Talking on the telephone can be a problem for hearing impaired persons. Today's society offers solutions like text- and video telephones, but to use these both parts need the special equipment, which can be expensive and complicated to use. There are also relay-services, serving as a link between the normal hearing and hearing impaired, but people might be unwilling to use such means due to privacy loss.

HörStöd[3] is a project proceeding at CTT in parallel to SYNFACE, but instead of the talking head a text appears on the display. This text is transcribed from the input speech signal and is not entirely correct so the results have not been satisfactory so far. Furthermore, reading a subtitle is not natural behaviour during conversion and it is probably easier to get confused by one incorrect letter than one viseme[4] similar to the one spoken (as with a talking head). Investigations and improvements are still being carried on.

LipCell is an Israeli project with the same basic idea as that of SYNFACE, i.e. a telephone for hearing impaired persons, but is not considered a strong competitor to SYNFACE in terms of quality. First of all, only the mouth is displayed in the LipCell project, so important information, such as eye-brow movements and nodding, is excluded. Le Goff *et al.* (1994) showed that this resulted in a lower intelligibility level than a face model when the audio is degraded. Moreover, the markers for turn taking (see 2.2.3) are rather primitive. Finally, the research behind this project does not seem as extensive as that of SYNFACE. Nevertheless, no thorough investigation has been realized so these intuitions might be incorrect.

[3] See homepage of CTT, http://www.speech.kth.se/ctt, for further information.

[4] A *viseme* is a facial image representing one or more phonemes. Term coined by Fisher 1968.

2.3 AUDIOVISUAL ASPECTS

2.3.1 Fundamentals

The multimodal aspect of speech explains the fact that speech can be produced and perceived through different senses: primarily, hearing and vision. In addition, touch can be an important channel for the hearing impaired persons (Bernstein & Benoit 1996; Oerlemans & Blamey 1998). This will, however, not be treated in this report.

Contributing to speech *production* are the vocal folds, the velum, the tongue, the lips, the jaw and the teeth. Audition and vision are *perceived* separately in the central nervous system, and the exact way in which they are integrated into a single audiovisual representation is not yet completely understood. Still, the results from various experiments measuring the electrical activity of the brain while perceiving audiovisual stimuli show that the integration between the two modes is most likely to occur in the dominant hemisphere, i.e. the left for most people.

As might be expected, but also according to Massaro (1980), the acoustic signal normally contains more linguistic information than the visual. Yet, in certain cases it is easier to see than to hear what is being said. Therefore, the focus may automatically alternate between the two modes, for instance when it comes to discrimination between front- and back consonants. Massaro & Cohen (1993) report that in an acoustically bad environment the distinction between the front consonants *b*, *m* and *p* could be problematic, since the lip movements are the same, i.e. they form a *viseme*. On the other hand, the back consonant *d* is easily visually separated from *b* despite the similar acoustic features.

Integration between the two modes of stimuli is likely to occur even when they should not match, e.g. when the face and voice do not represent the same sex or location (Massaro 1980), but also when there are discrepancies in the timing between sound and image. If the brain is unable to integrate the channels, the event will be perceived as coming from two different sources (Schomaker *et al.* 1995). Campbell & Dodd (1980) reported that the intelligibility for CVC-syllables exceeded the performance of the audio-only condition for audio delays up to 1600 ms. This should confirm that the interval for integration between the two modes is fairly wide, even though the ability to understand the message (i.e. the intelligibility) diminishes as the asynchrony increases. The big tolerance could be due to the material used in the study mentioned above: even though research on *sentence* material (McGrath & Summerfield 1985; Pandey *et al.* 1986; Grant & Greenberg 2001) has not examined delays as large as 1600 ms, there is a possibility that the threshold would be found for a lower value. Grant & Greenberg report for instance that an audio-leading delay of 400 (i.e. -400[5]) equalled the score for the audio-only condition. It is believed that when investigating simple speech sounds of short duration, such as syllables (CVC[6], VCV, CV and VC), the person could remember the leading stimulus when the second channel appears and hereby give a correct answer. This would be due to the fact that the two modes do not have a temporal overlap for larger asynchronies, which

[5] A delay with a negative sign denotes audio *leading* stimuli.

[6] C = consonant, V = vowel.

they do when it comes to sentence material: the channels are then perceived simultaneously which increases the risk for confusion. It is also likely that memorizing would be harder for sentences than for syllables.

Even though integration is possible for large asynchronies, the *detection* of conflicting data is possible for asynchronies as small as 79 ms for audio-leading (-79 ms) and 138 ms for audio-lagging (+138 ms) stimuli, as found by McGrath & Summerfield, when investigating CV-syllables. These numbers show a greater sensitivity for audio-leading than for audio-lagging stimuli, thereby inducing an earlier awareness of the asynchrony, and causing a decline in the intelligibility scores.

2.3.2 Leading or Lagging Audio

As explained in the preceding chapter, former studies[7], investigating the effect of asynchrony in speech perception, have shown a bigger tolerance for cases where the visual signal leads the acoustic. This is likely to be due to the more frequent exposure of audiovisual information in face-to-face communication where optics precedes acoustics. A well-known fact is that light travels faster than sound, such that as a consequence the visual reception of a distant source precedes the auditory (Massaro 1980).

How big is this tolerance? By measuring detection and/or intelligibility for various delays, it is possible to get an answer to this question. Further on, the main focus will be on the latter, i.e. the intelligibility. It is not always easy to make comparisons between studies since they all represent different material: sentences or syllables; audio leading or lagging; large or small asynchronies. Also, different types of acoustic distortion are employed. A few tendencies can, however, be seen in order to elucidate the magnitude of delays where a supplementary face loses its purpose.

Pandey *et al.* (1986) examined the understanding of sentences for delays up to +300 ms and for all of these the score was still better than the audio-only condition, but a significant decrease of scores, compared to the synchronous condition, was found for +180 ms (S/N=-10dB)[8] and above. The corresponding value for McGrath & Summerfield (1985) was somewhere between +80 and +160 ms: at the lower number there was no significant effect whereas the result for the higher approached the *visual*-only condition (they did not test *audio*-only). Smeele (1994) and Grant & Greenberg (2001) tested delays up to +400 and +280 ms, respectively, and all of these delays showed better scores than the respective audio-only condition. At the largest negative delay of Grant & Greenberg, i.e. -400 ms, the intelligibility equalled the audio-only condition, but already at -120 ms the intelligibility was as low as 32% (for -40 and -80 ms the score is about 46%). Smeele reports a significant decrease at -160 ms compared to synchrony. These results are consistent with the discussion

[7] Concerning intelligibility: Smeele 1994; Grant & Greenberg 2001. Concerning detection: Dixon & Spitz 1980; McGrath & Summerfield 1985.

[8] In analog and digital communications, signal-to-noise ratio, often written S/N or SNR, is a measure of signal strength relative to background noise. The ratio is usually measured in decibels (dB)." (http://searchnetworking.techtarget.com)

above concerning the fact that subjects react to smaller asynchronies for the audio-leading conditions.

Is perfect synchrony the ultimate case for perfect audiovisual perception? Certain studies show that it is not evident that this is the condition rendering maximum intelligibility. Whereas considerable reductions of the scores are found even for the shortest audio leading conditions results from audio-lagging experimental conditions show that a displacement of up to +80 ms have no significant effect on the intelligibility (among others Pandey *et al.* 1986 and McGrath & Summerfield 1985). According to the latter, a *gain* in intelligibility even occurred for delays between +60 ms and +120 ms. Exactly when this gain arose depended on the sound quality: S/N varied between 0 and -10 dB and the lower it was, the more the maximum score seemed to approach synchrony. Moreover, the top score in the study of Grant & Greenberg (2001) was at +120 ms and the performance did not drop significantly until the delay exceeded +200 ms. It is worth noticing that the discrepancies in these values, though small, could be due to the variations in sound distortion: Pandey *et al.* (1986) and McGrath & Summerfield (1985) made use of noise, whereas Grant & Greenberg (2001) employed a vocoder-like speech distortion technique (see 3.2.1).

Thus, several studies draw attention to this unnatural circumstance where audio-lagging delays give a better result than synchrony. The question is why? One reason could be that experimental uncertainties could have an influence but this is seldom discussed in the respective papers. The reason could also be as discussed in the first section of this chapter: people are more used to input where the optics precedes the acoustics.

2.3.3 Natural or Synthetic Face

Investigations which have examined the influence of synthetic faces on speech perception only examined the synchronous case. What can be concluded from these studies is that there is proof (Agelfors *et al.* 1998; Lidestam *et al.* 2001; Siciliano *et al.* 2002b) that even a synthetic face can serve as a helpful supplement to incomplete acoustic signals to most people, though to a somewhat lesser degree than for a natural face. This is probably due to the constant exposure of real faces in every person's everyday life. A synthetic face, on the other hand, might lack certain features which would affect the comprehension negatively. Lidestam *et al.* assumed that a natural face might have a better capacity to activate the so called multisensory neurons[9] than a synthetic face has. These neurons are thought to elicit responses that are far more powerful than the sum of those neurons treating each mode separately.

The studies mentioned above have been carried out on normal hearing as well as hearing impaired persons, with different materials (syllables and sentences, noise and vocoder-technique) and for different languages (Swedish, Dutch and English). This ought to demonstrate the broad usage of animated faces.

[9] These neurons were found by Barry Stein: http://www.wfubmc.edu/nba/faculty/stein/stein.html.

Many of the aspects discussed in the preceding chapter also concern the synthetic face. Therefore, no further explanation will be made here.

3 EXPERIMENT

Considering that the incoming speech controls the facial movements in the SYNFACE prototype, it is obvious that a certain delay must occur. Both the software and the hardware are responsible for this delay (Ward 2002). However, the task of this study is not to look at the causes but to determine the lower boundary for which the speech recognizer has enough time to provide an animation of quality and the higher acceptable boundary above which the talking head loses its purpose. For the experiments, this is done without the speech recognizer in order to avoid recognition errors. The potential quality of the synthetic face can thus be examined.

Instead of making these experiments exclusively for the SYNFACE application, a more general approach was chosen. This explains the use of the natural face as a complement to the synthetic as a visual mode, and also why not only audio-leading, but also audio-lagging asynchronies were examined. In this manner the natural face served to (1) examine the quality of the synthetic face relative to the natural, and (2) comparing the present study with previous ones for both audio leading and lagging delays. It was also important to point out the delays, if any, for which the intelligibility scores were found below the level of the audio-only condition.

Based on research discussed in chapter 2.3, the following outcome was expected in the present study:

1) better score for the natural face than for the synthetic
2) better score for the audio-lagging than for audio-leading experimental conditions
3) it is not obvious that the maximum score will be found for perfect synchrony
4) the intelligibility should deteriorate somewhere in the interval [0,-80] when the audio signal leads and [0,+240] when it lags

3.1 SUBJECTS

Seven female and five male subjects of 24 to 52 years of age (mean age 29, median 26) participated in the study. Seven of them were staff and students from KTH whereas the rest worked or studied elsewhere in Stockholm. They were all normal hearing and had normal or corrected-to-normal vision. All had Swedish as their mother tongue. One of them had performed similar tests before but that was not considered a problem for this study, since that was a long time ago. Another subject had some earlier experience from speech-reading. For their assistance the subjects received cinema tickets.

3.2 STIMULI

3.2.1 Audio

Since the subjects were normal hearing, the audio signal had to be treated in order to simulate hearing loss. This can be done in several ways. Some researchers have replaced the audio signal by rectangular pulse trains synchronised with the talker's glottal pulses. Others have used a full spectrum speech signal with additional noise, which can cause problems since fricatives consist of hissing sounds that are easily masked in noise. In the present study, a vocoder-like technique was used which is thought to present a signal similar to that provided by cochlear implants (Siciliano *et al.* 2002a). Within the region 100-5000 Hz three separate bands were chosen to provide the auditory information. Using this number of bands was believed to provide a good level of difficulty after observing the results from Siciliano *et al.* (2002b) who investigated the effect of both two- and three-band signals in synchrony. The remaining spectral regions (i.e. the bands) were then also excited by noise.

3.2.2 Face

The former study implemented by the Swedish SYNFACE group, included in the Siciliano *et al.* (2002b), used Sven (see 2.2.2) as the synthetic face. In order to make a comparison possible, the same choice was made this time. Consequently the type of face would not be a factor in any discrepancies between the two studies.

3.2.3 Sentences

The sentences were first developed by MacLeod & Summerfield in 1990 and then adapted to Swedish by Gunilla Öhngren (Dahlquist 2002). All sentences have a simple grammatical structure and approximately the same duration, and they are organised in 12 lists all of which represent the same level of difficulty. Examples of sentences, key words underlined:

- "<u>Fågeln</u> <u>bygger</u> ett litet <u>bo</u>" (The bird's building a nest)
- "<u>De</u> <u>plockade</u> några <u>hallon</u>" (They picked some raspberries)
- "<u>Äppelkakan</u> <u>var</u> för <u>varm</u>" (The apple pie was hot)
- "<u>Damen</u> <u>vattnade</u> <u>blommorna</u>" (The lady watered her plants)
- "<u>Mannen</u> <u>tvättade</u> <u>bilen</u>" (The husband cleaned the car)

In the present study, one test list consists of 14 sentences, produced and treated as described in the next chapter. In order to make the subjects accustomed to the faces and to the sound before the real tests began, a practise list was made from a double set of the remaining 12 sentences. This list contained all experimental conditions but ±300 ms (see 3.2.5).

3.2.4 Processing of Material

The original material consisted of video recorded sentences (*avi*-format) spoken by a Swedish man. By extracting the auditory signal, the desired disturbances (see 3.2.1) could be added and the different delays could be programmed. These new audiofiles (*wav*) replaced the original sounds in the *avi*-files. They also provided the auditory signals for the synthetic face.

The next step was to generate the individual coupling of test list and experimental condition. No subject had the same combination of these two components as anyone else. Randomization of the couplings, followed by conversion to *xml*-format, resulted in the final test material for each subject. The *xml*-files were then presented to the subjects by *Avplay*[10], which is a software developed by Jonas Beskow.

For more detailed information about video and sound, see Siciliano *et al.* 2002b!

3.2.5 Experimental Conditions

As discussed in chapter 2.3, the asynchrony can be fairly large before the intelligibility falls below the level of the audio-only condition. For the SYNFACE-application, however, it is also a matter of effective communication which demands small delays.

As a first step, the objective was to get a rough estimate of where the intelligibility would fall. The experimental delays were therefore distributed over a large interval ([-300, 300]) and with rather low resolution. If necessary, further testing would be made at a later stage. The audio-leading asynchronies were chosen for both faces whereas the audio-lagging asynchronies were used only for the natural face. This limitation is due to the small amount of material. Also, both faces were tested at synchrony in order to get a reference point.

Each combination of face and delay represents an experimental condition, for which the denotations are negative when the audio signal leads the visual and positive when it lags. The ones chosen for this study are shown as symbols in Table 1, where the dots signify the conditions considered the most difficult. One test list (14 sentences) was presented to each subject for each experimental condition.

Table 1. The experimental conditions. Each symbol signifies one experimental condition used in the present study. The dots show the conditions considered the most difficult.

Delay (ms)	-300	-175	-50	0	50	175	300
Natural face	•	•	×	×	×	×	•
Synthetic face	•	•	×	×			
Audio-only				•			

Every subject was given a personal set of conditions. The same sentences were used, but with different conditions for each subject (12 subjects, 12 experimental conditions). This was done to avoid any relation between a certain sentence and a specific modality. Furthermore, the different conditions were randomly ordered so

[10] http://www.speech.kth.se/~beskow/avplay/avplay.html

that each subject was given an individual sequence. In this manner the sequence would not influence the total score: a condition with a large delay, e.g. *natural face -175*, following *natural face -50*, could otherwise render a higher score than reasonable. The notation for a specific subject will, in this report, be S*x* where *x* represents the subject's number.

3.3 PROCEDURE

The experimenter ran the test through remote-controlled device (see Figure 3) and noted the subject's answer directly on the remote computer. The log-files were immediately created in the experimental computer. A second screen was connected to the experimental computer to serve as a control as the test proceeded.

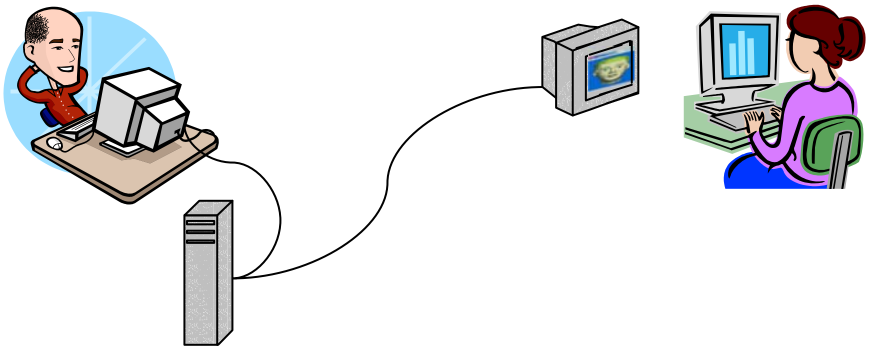


Figure 3. *Experimental set-up. The subject is seated to the left, in front of the experimental computer. The experimenter runs the test from the remote computer, but can also observe the stimuli through the control monitor.*

The only environmental requirement for these experiments is silence to avoid any added disturbances. The trials were therefore implemented in a sound treated room where the subject was seated in front of a display that would present the three different visual stimuli (the natural and synthetic faces as well as the loudspeaker). The resolution of the video was 720x576 and the frame rate 25 Hz. The synthetic face was shown in approximately the same size as the natural. The audio signal was sampled at 32 kHz and played back through headphones in order to block any surrounding sounds and to facilitate the focus on the stimulus. If necessary, the subject was allowed to adjust the volume before the test started.

The test material was presented at two sessions. Session I included instructions, practise trials and the first 6 conditions and lasted between 30 and 50 minutes. The span of session II was 20 to 35 minutes and included the remaining 6 conditions as well as a questionnaire. No subject had more than four difficult (see Table 1) experimental conditions in one session, as could be seen by checking the outcome of the randomization. This should prevent any feeling of failure. Although the aim was to run the sessions on two consecutive days, this was not the case for two of the subjects: a weekend passed between the two occasions for S7 while for S12 only half a day. This was, however, not considered a problem for the outcome of the tests.

Prior to the first session the subjects had received instructions in Swedish about the purpose of the study (see Appendix 1). At the start of the experiment they also learnt about the experimental conditions (but not the sequence) they would experience and that the sound would be very poor. They were encouraged to guess, even though what they thought they heard would not make sense.

In connection to the practise list, a direct feedback was given to the subject after each sentence to let him/her become more accustomed to the sound and to the structure of the sentences. During the real test, no feedback was provided until after completion of the tests. For each condition, the 14 sentences were presented to the subject who was instructed to say out loud what he/she perceived. To avoid confusion the subject was informed when a new condition would start.

3.4 CORRECTION OF LOG-FILES

The first two sentences for every test list were regarded as practise material and therefore discarded in the scoring. With 12 sentences of 3 key words, a total of 36 key words per subject could be achieved for every experimental condition. The intelligibility was then obtained by determining the percentage of this total.

If the root of a key word was correct, 1 point was given, regardless of inflexional form. Sometimes the subject would understand the surrounding words and exchange one key word for another word with the same meaning, e.g. "åkte" and "reste". During a conversation in real life this substitution would be acceptable, but since the subject was instructed to answer exactly what he/she perceived, this would disagree with the rest of the test and, therefore, no point was given. A switch between "han" and "hon" might seem like a mild error, especially since the user, in any communication, most likely would know who the conversation is about, but in order to be consistent no point was given. Furthermore, no point was given for a compound key word when only half of the word was correct, as when answering "rullstolen" or "stolen" instead of "gungstolen".

4 RESULTS AND ANALYSIS

The intelligibility scores for the entire experiment are shown in Figure 4. Later on, the different elements of the experiment will be examined in detail. As can be seen, the expectations made at the beginning of chapter 3 were correct to a high degree: the score for the natural face is somewhat better than for the synthetic at all conditions but one; the intelligibility is a bit higher for the audio-lagging delays than for the leading; the maximum score is not found at synchrony but for +50 ms (natural face); the intelligibility seems to drop between the -50 ms and -175 ms conditions for the audio-leading delays and between +175 ms and +300 ms the for audio-lagging. Important is also the fact that no bimodal condition had a score below the audio-only condition. The significance in these findings will be discussed later in this chapter and in chapter 5.

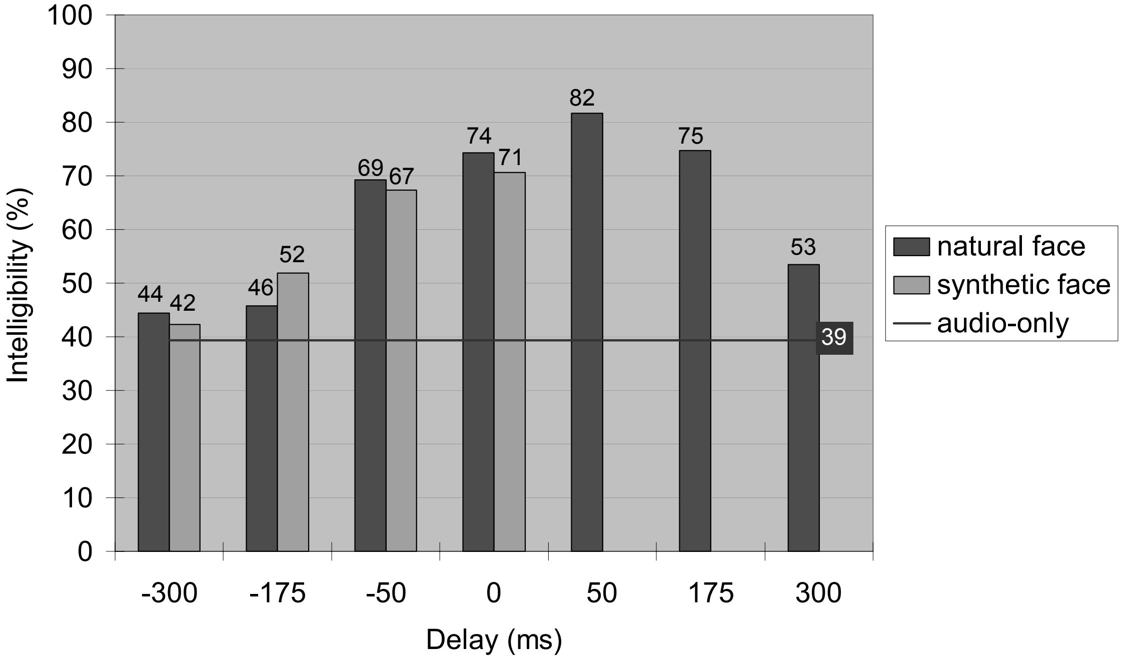


Figure 4. Total Score*. Columns show the intelligibility for every experimental condition summarized for all subjects: the dark columns are for the natural face and the bright for synthetic face. The horizontal line shows the total intelligibility for the audio-only condition.*

Figure 5 and 6 present the results for the two visual modes separately. The bars, representing ±1 standard deviations, reveal substantial variation between subjects. The audio-only condition, which can be observed in Figure 4, has a standard deviation of about the same size: ±24 percentage points.

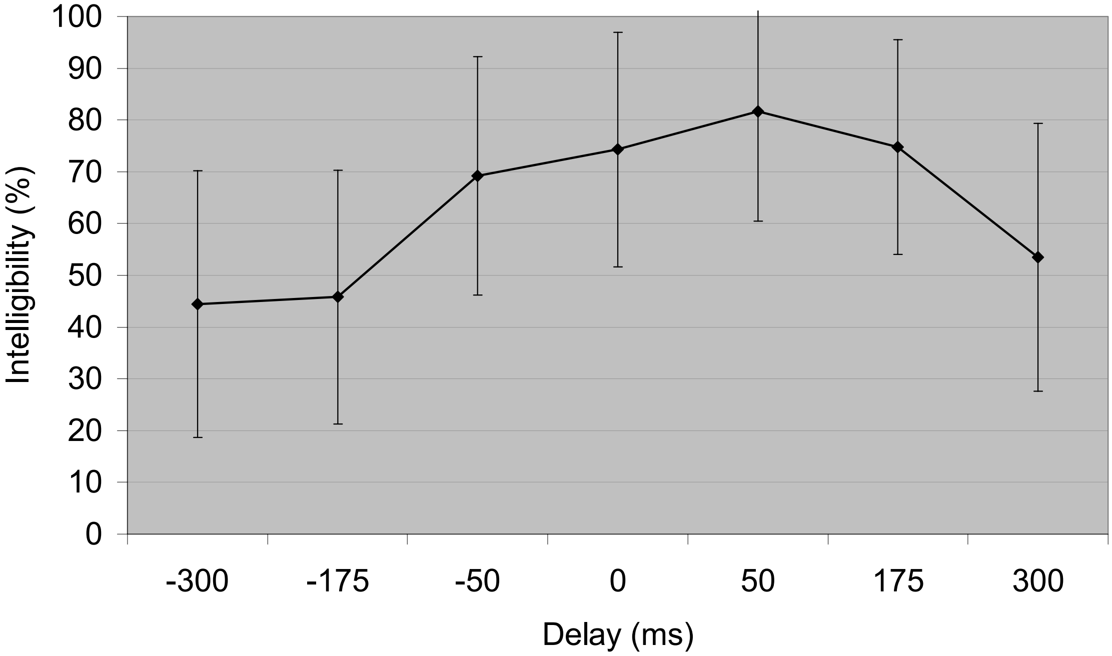


Figure 5. *Total score for the natural face displayed by a line. Bars represent ±1 standard deviations of the means.*

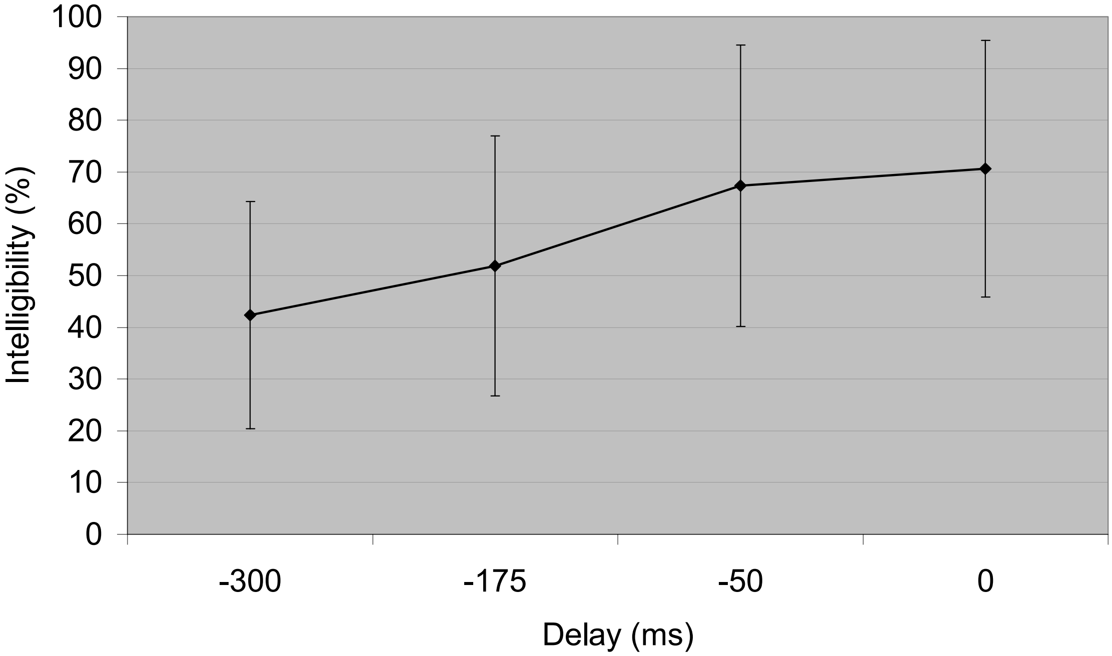


Figure 6. *Total score for the synthetic face displayed by a line. Bars represent ±1 standard deviations of the means.*

4.1 LEADING vs. LAGGING

The natural face is the only visual mode that allows comparison of the leading and lagging auditory signals. In Figure 4 and 5, the delay that renders maximum intelligibility seems to be located on the right side of synchrony, i.e. for audio-lagging delays. This could be due to experimental uncertainties which are further discussed in chapter 4.7.

A clear picture of the experiment can be obtained just by examining the charts, but through statistical analysis certain aspects can be elucidated. An ANOVA[11] with the intelligibility scores for the natural face as dependent and the amount of delay (7 levels) as independent variable revealed a significant effect of the delay on the intelligibility scores: [F(6,77)=5,0; p=0,0002] (see also Appendix 2, Table I). A Fisher's PLSD post hoc test (5% significance level) on the effect of the delay further showed significant differences between -175 ms and -50 ms, as well as between +175 ms and +300 ms, which therefore reveals the critic intervals for when the intelligibility deteriorate (see also Appendix 2, Table II). Even though the shape of the curve is almost as expected for [-50, +175], the differences are not big enough to show significance. Subsequently, delays within this interval should not matter for the understanding of a spoken message.

A comparison between leading and lagging audio signals is done in Figure 7 and 8 where the sum of the three leading delays was compared to the sum of the three lagging. The dots in the former show the *individual* scores for both cases. The line is drawn for *y=x* and is therefore *not* a regression line for the data but represents where the dots would be found if the score for both cases were identical. As can be seen, all dots are located closer to the lagging case. Even Figure 8, which presents the *total* score for each case, reveals an advantage for audio-lagging stimuli. If the line would be completely horizontal, there would be no difference between leading and lagging stimuli. Further evidence of the audio-lagging advantage is that 57% of all correct answers (of comparable delays) were for this category. ANOVA reveals that the difference between leading and lagging is significant: [F(1,70)=7,6;p=0,007], (see also Appendix 2, Table III).

[11] ANOVA = "Analysis of Variance, a statistical test for heterogeneity of means by analysis of group variances" (http://mathworld.wolfram.com/ANOVA.html). An alpha level of 0,05 was used for all statistical tests.

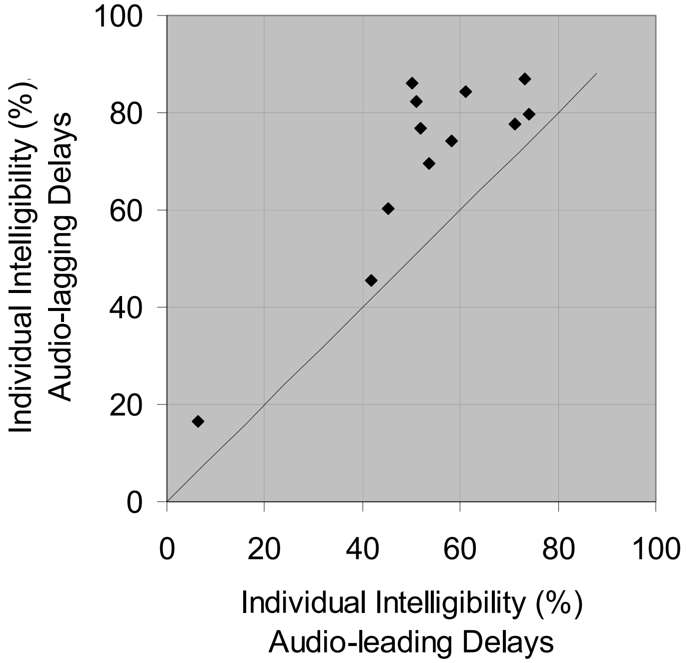


Figure 7. *Comparison between audio-leading and -lagging delays. Every dot represents the score for one subject. The closer the dots are to the line y=x, the more similar are the two modes.*

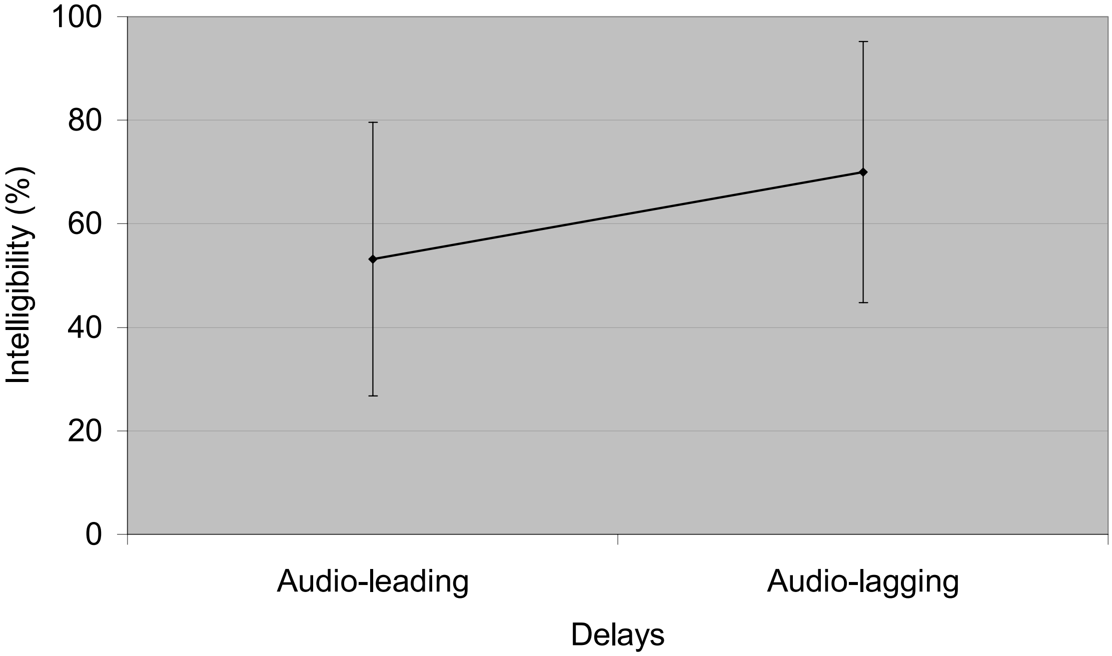


Figure 8. *Comparison between audio-leading and -lagging delays. The values represent the sum for all subjects. The more horizontal the line, the smaller is the difference between the cases.*

It is also interesting to compare a certain audio-leading delay with the corresponding audio-lagging. This is done in Figure 9 where a perfectly horizontal line would signify equal intelligibility between the two cases. The above mentioned Fisher's PLSD post hoc test shows that the differences between -300 ms and +300 ms as well as -50 ms and 50 ms are non-significant whereas -175 ms and +175 ms are significant, which also can be noticed just by watching the chart. The difference between the delays of -175 ms and +300 ms is non-significant, which implies that the intelligibility levels of these two experimental conditions are similar. This is further evidence that it is easier to understand stimuli where optics precedes acoustics.

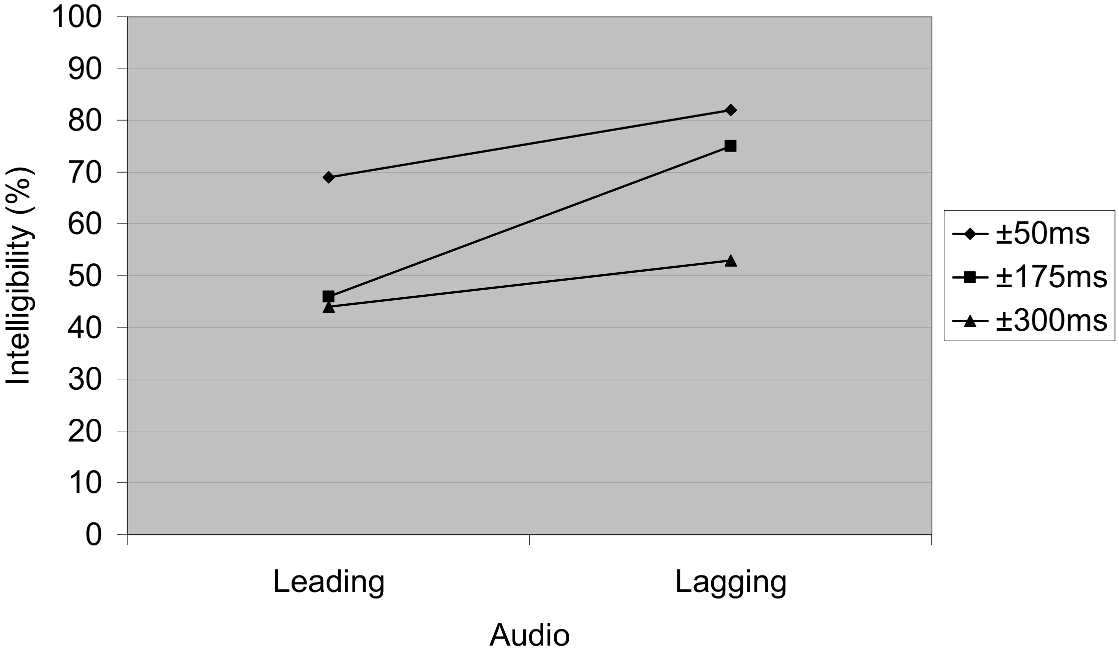


Figure 9. *Comparison with a certain audio leading delay with the corresponding audio lagging delay. The more horizontal the line, the smaller is the difference between the cases. The delay of ±175 ms stands for the biggest difference between leading and lagging stimuli.*

4.2 NATURAL vs. SYNTHETIC

For comparison between the two faces, the audio-lagging conditions must be discarded. As can be seen in Figure 4, the natural face shows better scores for all comparable delays but -175 ms. Standard deviations for the two cases are shown in Figure 5 and 6.

Figure 10 compares the individual scores for each face with the equality line, *y=x*, which serves as a reference for when there is no difference between the visual stimuli. Figure 11 represents the same feature but with the scores for all subjects summarized. Both charts show a minor effect of face.

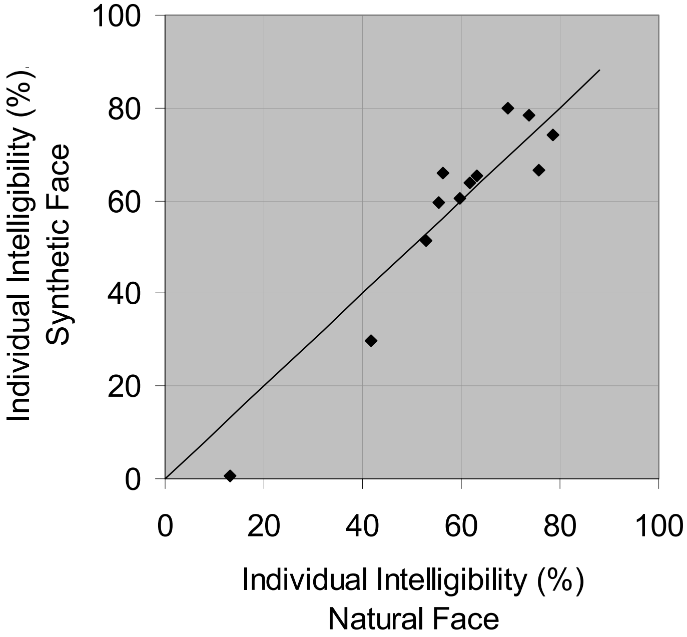


Figure 10. *Comparison between the faces. Every dot represents the score for one subject. The closer the dots are to the line y=x, the more similar are the two modes.*

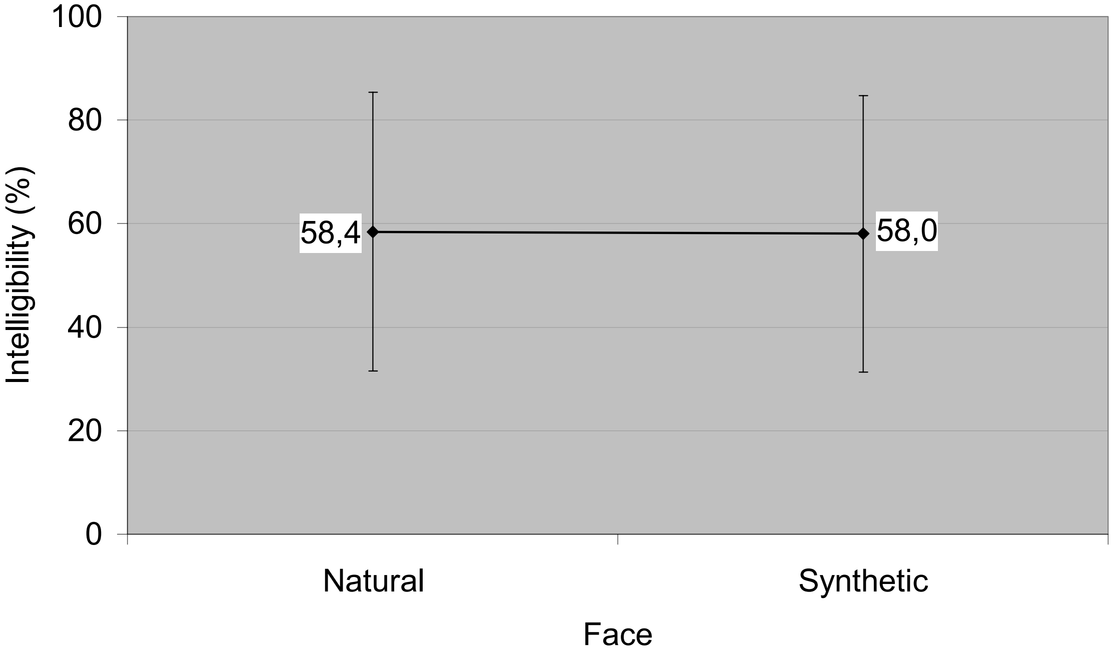


Figure 11. *Comparison between the faces. The values are shown as a sum for all subjects. The more horizontal the line, the smaller is the difference between the cases. The results are practically identical. Bars represent ±1 standard deviations of the means.*

In Figure 12 each negative delay is represented separately in order to get a more detailed picture of where the differences are found. The more horizontal the line, the smaller is the difference between the faces for that specific delay. It is clear that the delay of -175 ms is the one responsible for the equalization seen in Figure 11. However, no delay induces any big differences, and an ANOVA with intelligibility scores for the audio-leading conditions as dependent variable and the face and delay independent variables, verifies that the effect of face is non-significant: [F(1,94)=0,005; p=0,94], (see also Appendix 2, Table IV). Moreover, the natural face stands for 50.2% of all correct answers for these delays, which also indicates similarity between the faces.

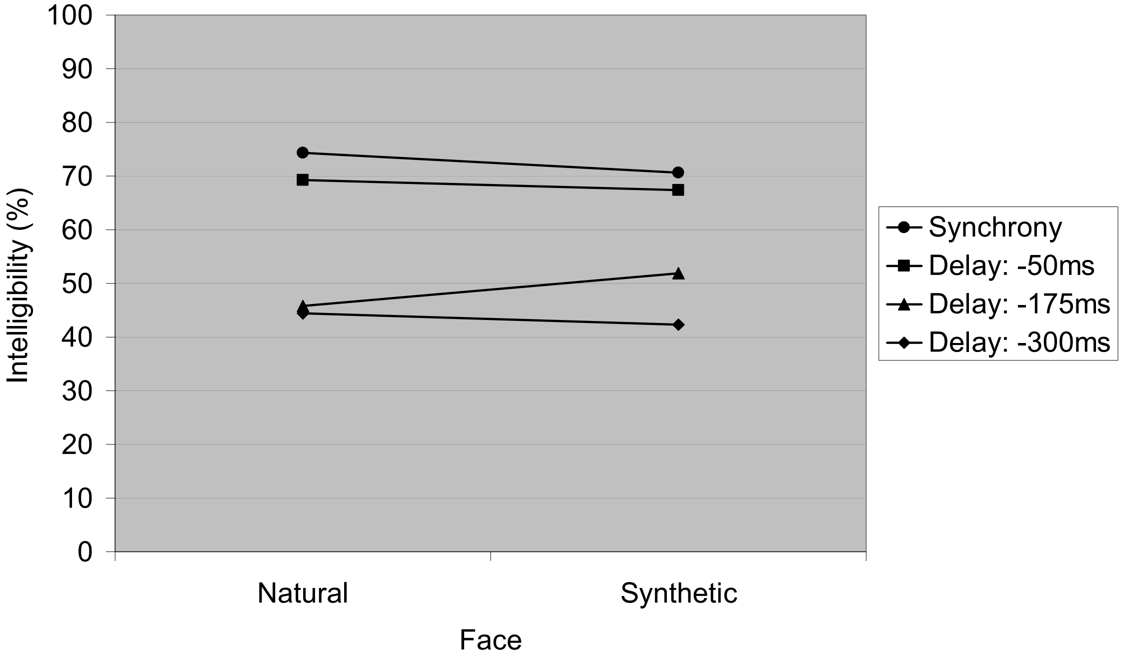


Figure 12. *Comparison between faces for each delay. The more horizontal each line, the smaller the difference between the faces for that delay.*

Is the advantage for the synthetic face at -175 ms only a coincidence or do the faces really represent an overall resemblance in scores? A check of at which positions *natural face -175* and *synthetic face -175* were found in the individual test list showed that the distribution across subjects was similar, i.e. no possible learning effects should matter for the results of these two experimental conditions.

4.3 AUDIO-ONLY

Even though no column is *below* the line for the audio-only condition in Figure 4, the columns representing the most difficult experimental conditions (see Table 1) are quite *close* to it. Accordingly, there might be a risk that the bimodal stimuli are not significantly better than the audio-only condition for all delays tested. A one-way

ANOVA, comparing the two faces with the audio-only condition separately, rightly shows that this is the case. So again, we can only be sure of that the face helps in the interval [-50, +175].

4.4 SEQUENCE OF STIMULI

To investigate a possible learning effect, the *first* test lists for all subjects were assembled, and then the *second* test lists and so on. Results are plotted in Figure 13. Test lists 1-6 in Figure 13 correspond to Session I and cases 7-12 to Session II. The value for the first test list is lower than the subsequent values. Therefore, regardless of experimental condition, the first stimulus seems to have had the function of habituation. After the first stimulus the performance improved. This suggests that the practice list should have been longer.

As can be seen, there was a small reduction in scores between test list number 6 and 7, i.e. between the two sessions. This might imply that a practise session should have preceded even the second test session. However, considering the small amount of subjects and material, it is essential not to take too much importance in these findings, especially since there are unpredicted reductions for test lists 10 and 11. There is a possibility that the picture would be more accurate with more subjects than 12, but further investigation is not necessary for the present study.

Even though Figure 13 shows improvements as the test proceeded, an ANOVA did not reveal any significance for the sequence of stimuli: [F(11,132)=1,2; p=0,32], (see also Appendix 2, Table V). This is probably due to the big variation in scores between subjects.

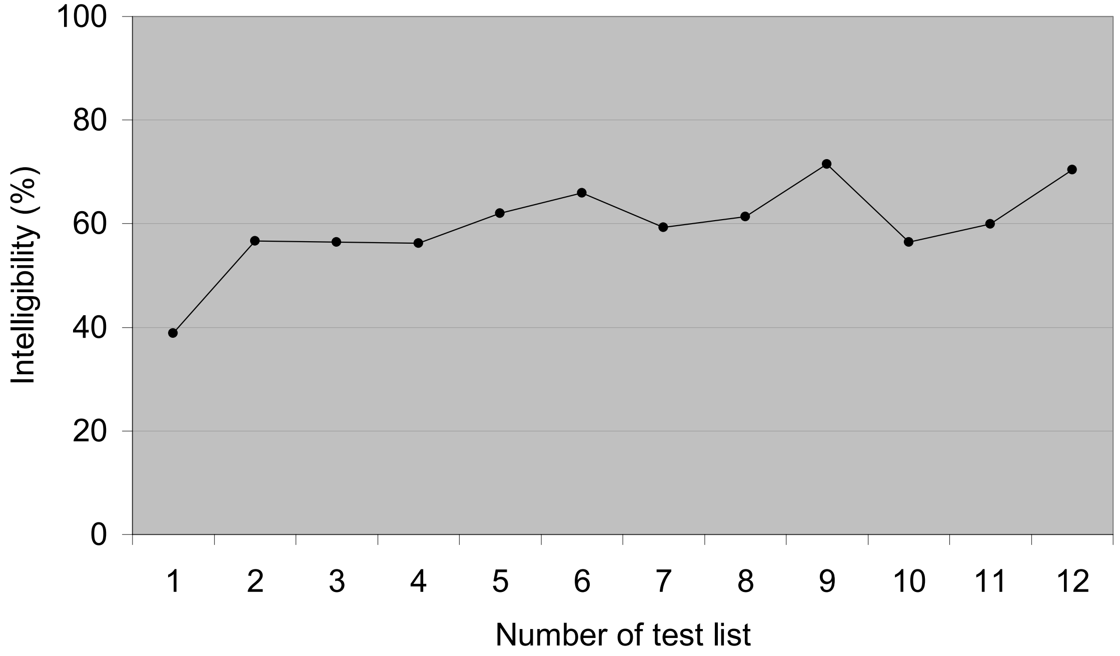


Figure 13. *Effect of the sequence of test lists. The improvement in scores is fairly large but has no significance statistically.*

Only one subject was better in session I than in session II but the scores for most of the remaining subjects were, however, still close to the equality line in Figure 14. A comparison between the two sessions reveals that 53% of all correct answers were in Session II, revealing that the difference is small. The sum of individual scores is presented for each session in Figure 15: perfect horizontality would indicate equal sums. These two charts suggest that some learning might have taken place, but, as stated above, this effect is non-significant.

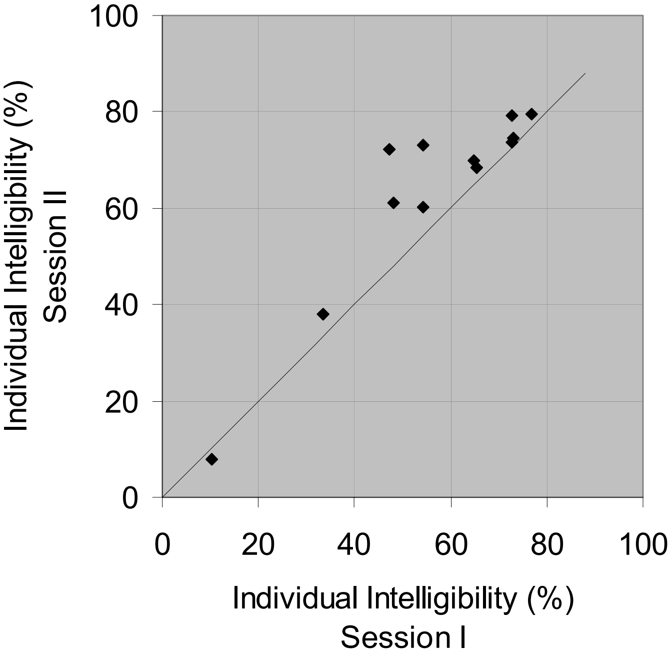


Figure 14. *Comparison between sessions. Every dot represents the score for one subject. The closer the dots are the line, y=x, the more similar are the two modes.*

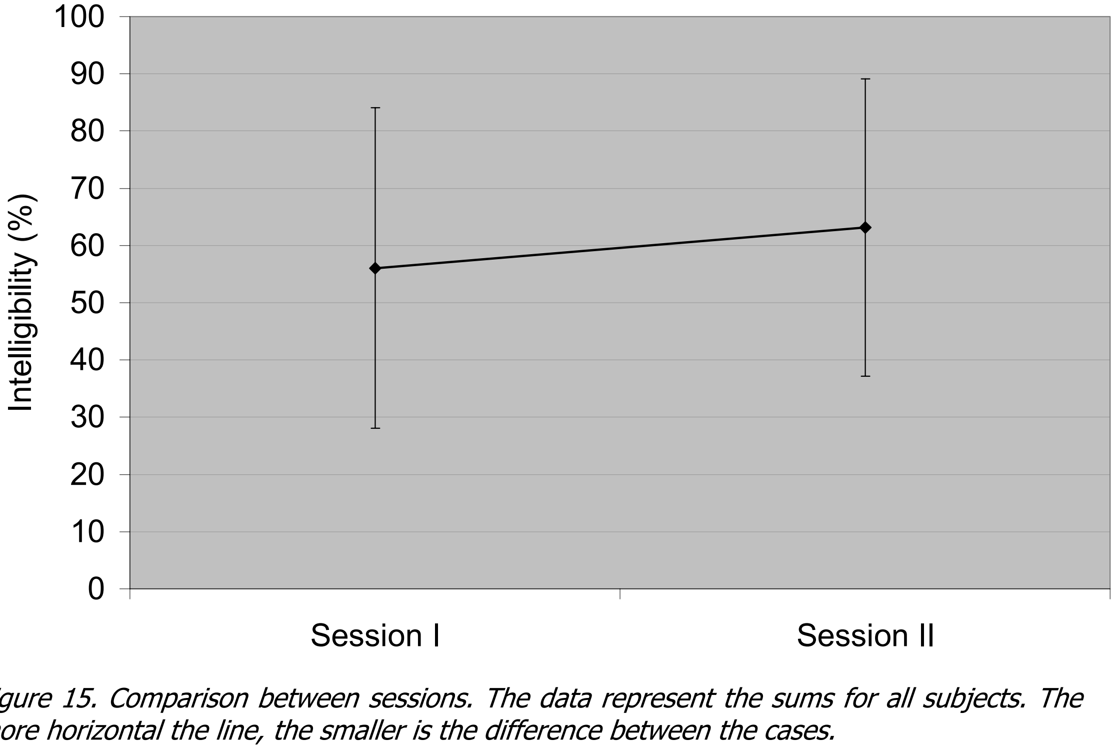


Figure 15. Comparison between sessions. The data represent the sums for all subjects. The more horizontal the line, the smaller is the difference between the cases.

4.5 LINGUISTIC MATERIAL

The difficulty of the different lists of sentences was intended to be equal as explained in 3.2.3. This was verified by an ANOVA, based on the results of the present study. A variation of about 20 percentage points between the separate lists can be seen in Figure 16. One reason for this could be that each list was represented by a different experimental condition for each subject. These couplings were randomized, as described in 3.2.4, but still, a greater number of participants would probably even out the slope in the chart. Despite the fairly big variation, the effect of list was non-significant: [F(11,132)=0,54; p=0,87], (see also Appendix 2, Table VI).

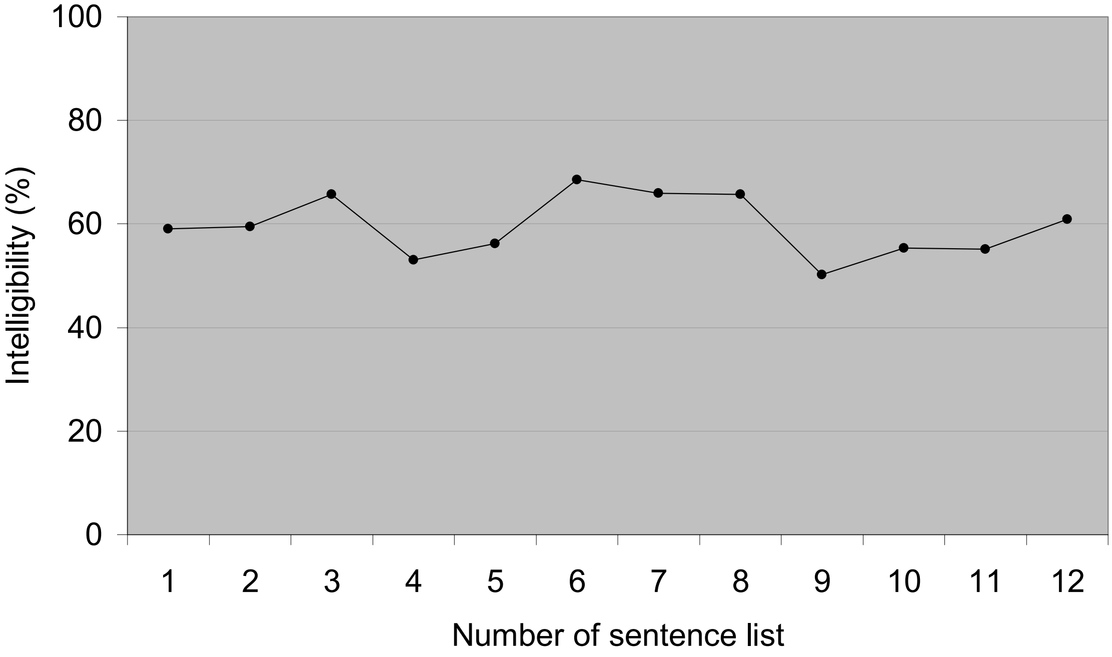


Figure 16. *The results for each sentence list across all subjects.*

4.6 INDIVIDUAL ASPECTS

Too much importance can not be attached on individual scores since the sequence, in which the experimental conditions were presented to the subject, can influence the individual results, as discussed in 3.2.5. Still, it could be interesting to elucidate some variations between subjects. Three examples are shown below.

S7 was the one who achieved the highest overall score and his results are shown in Figure 17. The columns in the chart reveal that he did not seem very affected by asynchronies. Noteworthy is that he, after the two sessions, explained that he rather focused on the face's forehead instead of the mouth when the time difference between the audio and the face got too disturbing. This was not the intention of the study, but the fact that he still had high scores by practically just listening, implies that he might not need a supplementary face to understand what is said. The same conclusion can be made by observing his score for the audio-only condition, which is fairly high.

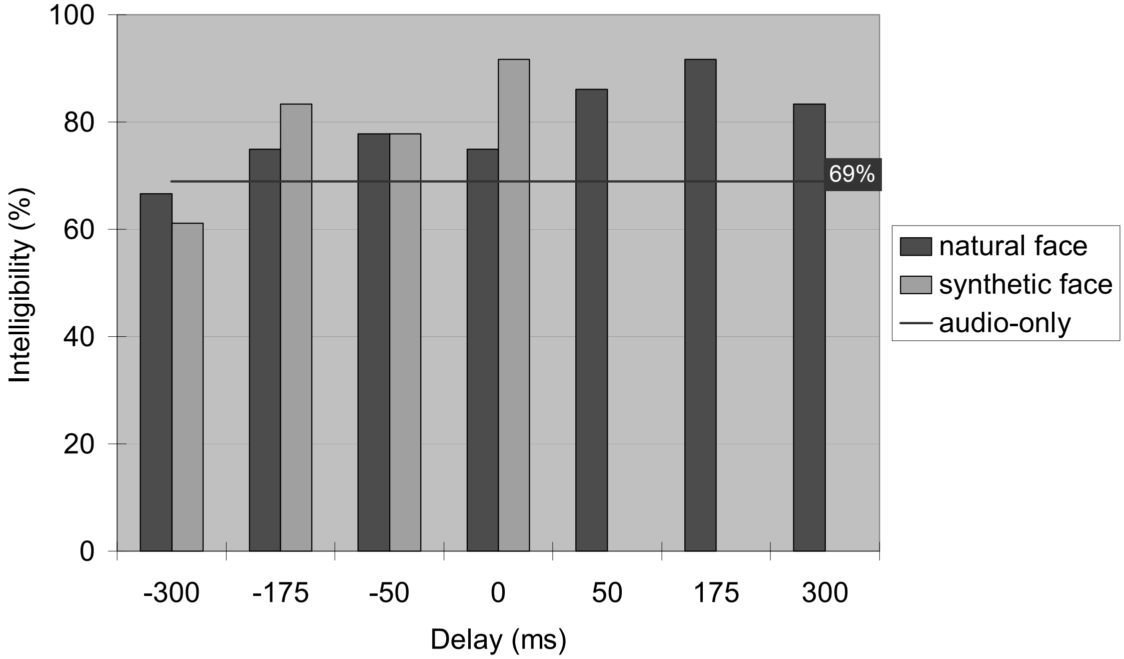


Figure 17. *Individual score - Subject 7. The line and the black box represent the audio-only score.*

S1 is found in the other extreme, see Figure 18. He had low values for all experimental conditions. His *highest* value was 33% (natural face in synchrony), which is nearly half of S7's *lowest* score, as can be seen by comparing the individual charts.

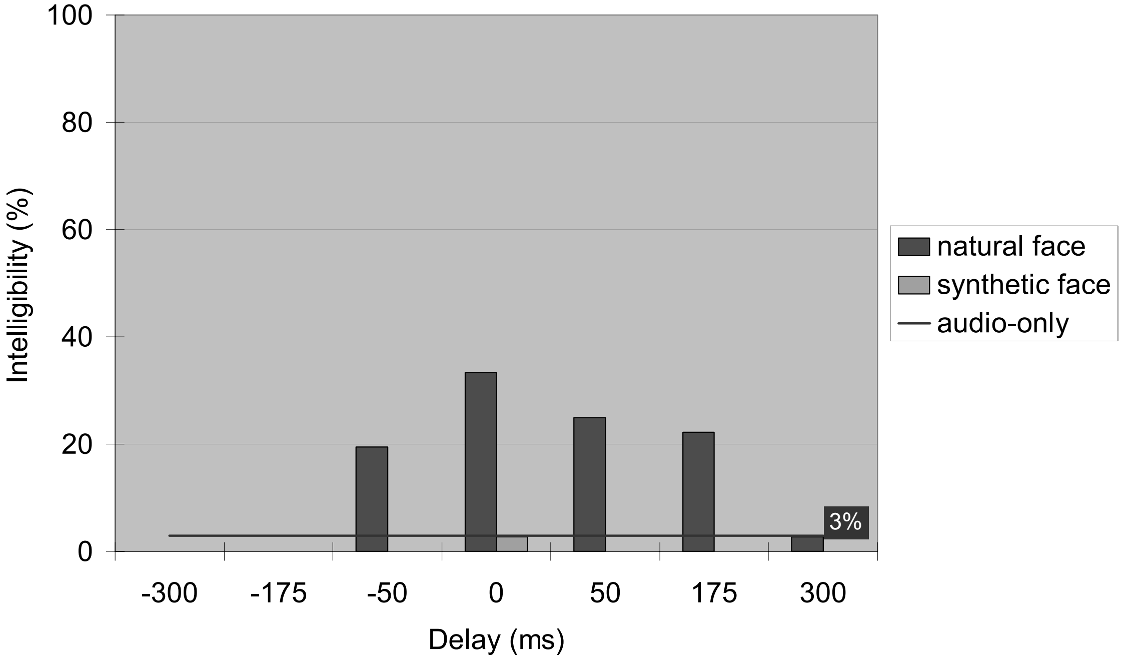


Figure 18. *Individual score - Subject 1. The line and the black box represent the audio-only score.*

S10 achieved better scores than S7 for the experimental conditions near synchrony, at least for the natural face, but unlike S7 she was very affected by large delays (see Figure 19). As can be seen, there is one experimental condition with a considerably lower score than for the remaining conditions for this subject, i.e. *natural face -300*. A check of her individual sequence (the table beside her individual chart in Appendix 3) reveals that this was her very first test list and could therefore function as habituation material, thus affecting her performance negatively. More discussion about possible learning effects is done in chapter 4.4.

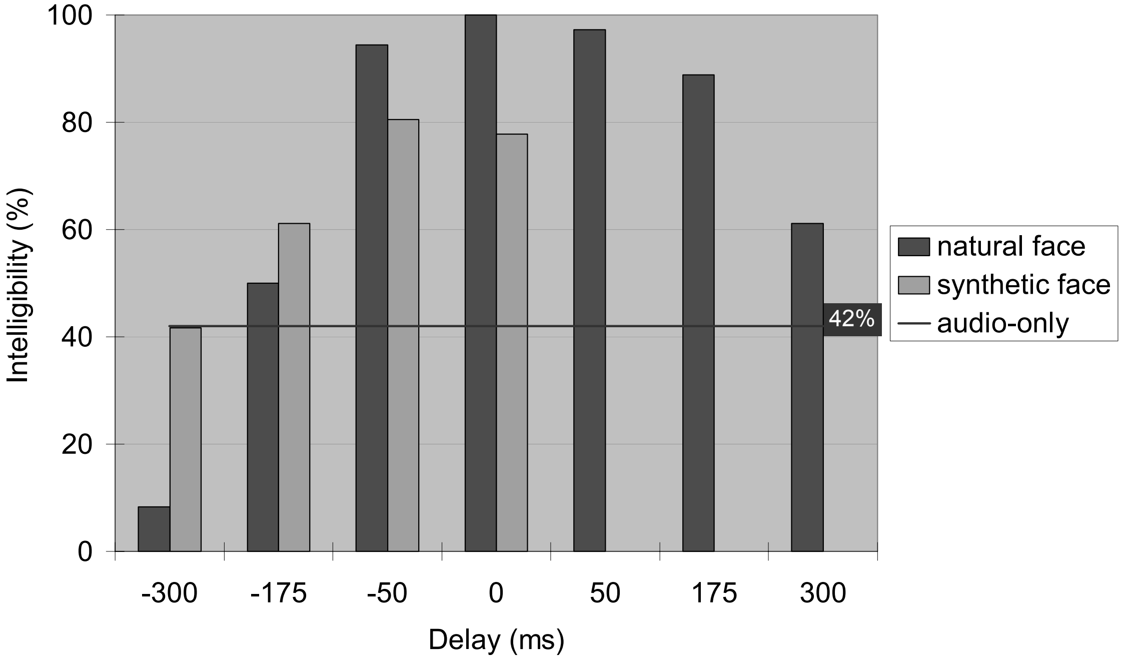


Figure 19. *Individual score - Subject 10. The line represents the audio-only score, which value is marked in the black box.*

The charts for the remaining subjects are shown in Appendix 3. Most subjects scored better than the audio-only condition for most of their bimodal stimuli. Exceptions could be due to the sequence of stimuli (habituation, learning). S5, for example, scored above the audio-only value for all experimental conditions but the ones at position 1 and 7 in her/his personal sequence, i.e. the first stimulus for each session. Other factors could also matter.

4.7 POSSIBLE SOURCES OF ERROR

In the preceding chapters a quite thorough investigation about the delay effect has been made. But are the programmed delay values reliable? Considering that +50 ms was the delay rendering maximum intelligibility scores in the tests, it might be suspected that the actual synchronous condition instead is displaced towards the audio-lagging case. Below, the size of possible uncertainties is discussed.

The errors, if any, of the soundcard and the update of the computer screen are below 10 ms and 14 ms (about 70 frames/second), respectively. The video is presented with a frequency of 25 frames per second, as described in chapter 3.3. This corresponds to a possible error of 40 ms, which is a magnitude approaching the smallest delay tested, i.e. ±50 ms. It is, therefore, not worth testing shorter delays than 40 ms. The update of the synthetic face is at least as good as the video.

4.8 THOUGHTS ABOUT THE PROCEDURE

4.8.1 The author's views

It was interesting to see the differences between the subjects: the way they prepared; the way they performed the tests; the way they succeeded. Regardless of the actual score, they seemed affected by the stimuli in different ways.

The ability to guess seemed to vary between the subjects. S1, for example, did not guess at all, but when he answered, it was often correct. He was the one with the lowest scores (see chapter 4.6). S8, on the other hand, took wild guesses of which many were wrong, but she still scored much better than S1. The results of S8 can be observed in Appendix 3. Others guessed to some extent, which in most cases resulted in correct or partly correct answers. It is possible that one reason for not guessing is fear of failure.

The response time also varied between the subjects: the individual test sessions were more time-consuming for the subjects who were considering or guessing what to answer, whereas the sessions for the ones not guessing, or giving immediate answers, were shorter in time. For the time span of the each session, see chapter 3.3.

4.8.2 The subjects' views

As mentioned in chapter 3.3, all subjects filled in a questionnaire after completion of both sessions. Here, a few of their thoughts are presented.

One subject (S4) expressed a slight irritation of obsolete gender roles found in the sentences. Apart from the irritation factor, this could also affect the results by facilitating the guessing when similar sounds ("han" ⇔ ″hon″ and "mannen" ⇔ ″damen″). S7 said that he easily saw patterns in the structure of the sentences, which also made the guessing part easier.

Four of the subjects (S1, S8, S10 and S11) expressed a difficulty to speech-read the synthetic face because of its rigid movements but when observing the actual scoring this was obvious only for S1.

S7 reported that he only used the lips from the visual channel, and that it would be better to exclude the rest of the face on order to direct the focus to the mouth (this subject was the one with the highest overall score, see chapter 4.6). This is in opposition to what was stated in 2.2.4: the mouth is not as informative as the entire face is. Also, others said that they did use other parameters of the face in order to comprehend the message: eye-brow and head movements, for example, sometimes helped separating two words from each other when the stimuli represented synchrony or small delays. In other words, those supplementary motions helped mediating the prosody.

The bad sound first shocked the subjects and they thought there was something wrong with the stimuli. They were, therefore, surprised and encouraged by the improvement in comprehension as the test proceeded. S1, S2 and S5 experienced a variation in audio quality, even within one experimental condition. When listening to the sentences without watching the face, I got the same impression myself: the audio sometimes seemed to vary in intensity. Still, I do not consider this being a factor contributing to any discrepancies in the results.

All subjects found the test interesting and amusing, and most of them thought that SYNFACE would be a helpful product for hearing impaired persons.

4.8.3 What makes a person good at this kind of experiments?

It seems like a certain level of personal maturity encourages guessing and avoids fear of failure. The ability to focus on the stimuli and to have a positive attitude about coming stimuli (instead of being disappointed about failing the preceding sentence) is also of importance. A competitive nature sometimes seemed to block a person during the tests: the desire to get a high score made him/her hesitant to guess.

S12 believed that a good ear for music could help a person to hear the message. He said that he, during the test, could remember and then retrieve the intonation to construct the answer. This was one of the subjects who achieved high scores (see chapter 4.6 and Appendix 3). To investigate this possible factor, an e-mail was sent to all subjects asking them to rate their ear for music on a scale with the steps 1, 2, 3, 4, where 1 would signify a poor musical ear and 4 a good musical ear. When comparing each subject’s rating with the respective test results, a possible relation was slightly visible for some subjects, but not for all.

It might be expected that persons with earlier experience of speech-reading would be more successful than others in these experiments, at least for experimental conditions around synchrony. It is then reasonable to believe that they instead could be more distressed by large asynchronies, since they probably rely more on the visual mode than others. McGrath & Summerfield (1985) divided the subjects into three sub-groups (poor, average and good speech-readers) and the more speech-reading skills they had, the more affected they seemed by asynchronies: the group without experience had a significant decrease in scores at +160 ms whereas the threshold for the rest was at +80 ms. Pandey *et al.* (1986), on the other hand, stated that the main effects were similar between all subjects except in the visual-mode, where the experienced speech-readers achieved higher scores. In the present study, only S4 and S5 informed about *some* earlier experience, but no patterns can be confirmed statistically.

In summary, there seem to be several factors affecting the results and every subject seems to be influenced by different combinations and proportions of these factors. What do these conclusions mean for the future use of a product like SYNFACE? Not much: for the target group of SYNFACE the competition factor is gone; fear of failure is likely to be reduced; and the user should be highly motivated to use the product.

5 DISCUSSION

The purpose of the present study was to examine how the perception of bimodal stimuli is affected by asynchronies as well as by different types of faces. The outcome was intended to clarify if a talking head can be an aid for hearing impaired persons in telephone communication and also to serve as a reference for similar studies investigating the perception of audio-leading and -lagging stimuli.

The expectations made at the beginning of chapter 3 appear to be true, though with a few exceptions and without considering the standard deviations, i.e. the bars in the charts.

1) A quick glance at the diagram presenting the total scores (Figure 4) might give the impression that the natural face represents an intelligibility level that can not be fully reached by the synthetic face. However, when investigating the results in further detail, it is clear that the difference between the faces is not very big, especially since the delay of -175 ms provided even better score for the synthetic face than for the natural. Figure 10 and 11 shows similar scores for the faces and an ANOVA, including all of the comparable conditions (i.e. the audio-leading as well as the synchronous), reveals that the effect of face is non-significant.
2) An instantly recognizable advantage for audio-lagging signals can be observed in Figure 7: all subjects achieved higher scores for this case, regardless of the level of their overall performance. ANOVA shows significance for the effect of preceding channel. Also, when comparing a specific delay with the corresponding case on the other side of synchrony, all delays show better scores for the lagging stimuli. It is, however, only the difference between -175 ms and +175 ms that is significant.
3) By examining Figure 5, audio-lagging stimuli seem advantageous compared to synchrony. As discussed in 2.3.2, this could be due to experimental uncertainties and/or the greater experience of optic-leading stimuli in real life.
4) It was expected that the intelligibility would degenerate before -80 ms for stimuli where the audio signal precedes the visual. From the delays tested in the present study, it is clear that a considerable deterioration occurs between -50 ms and -175 ms, but it is impossible to know the exact position without a greater resolution of delays. Moreover, ANOVA shows significance in the difference between these two delays. A maximum threshold of +240 ms was expected for audio-lagging stimuli and this also seems to be verified by analysis of the results: ANOVA revealed a significant decline in between +175 ms and +300 ms. It would be preferable to investigate asynchronies between the ones tested, in order to get a better picture of what happens and where it happens.

Also, as explained in chapter 2.3, the intelligibility dropped to the audio-only level at -400 ms for Grant & Greenberg (2001). In the present study the score approached this level at -300 ms so the results are similar.

One objective with the present study was to compare it with previous research and also to serve as a reference to coming projects. The results found here can easily be compared with earlier research by observing chapter 2.3 of this report. The only study truly comparable with the present study is the Swedish part of Siciliano *et al.* (2002b), i.e. the study also performed within the SYNFACE project. The results shown in Table 2 were attained by employment of the same faces (natural and synthetic), the same audio distortion (vocoder with 3 bands) and the same material (language and sentences). The higher scores for the present study should therefore be due to other factors, such as the selection of subjects or differences, though small, in the experimental procedures.

Table 2. Comparison of the intelligibility levels in the present study and Siciliano *et al.* (2002b). Only the results representing features that are common to the two studies are presented (face, audio distortion and language).

	Siciliano *et al.* (2002b)	Present Study
Audio-only	32.5%	39.4%
Natural Face, synchrony	66.0%	74.3%
Synthetic Face, synchrony	60.7%	70.6%

5.1 FUTURE WORK

In the present study, isolated sentences were presented, which could complicate the understanding of the message. For the future use of SYNFACE this should be avoided, since in any communication (face-to-face or via telephone), a specific topic is discussed which makes it easier to expect certain words or phrases. This is supported by Lidestam *et al.* (2001), who included an investigation of contextual cueing in their experiments, which concluded that topic-specific information provided during the tests increased the scores.

As Ward (2002) concluded, SYNFACE could be used as it is today but the delay would most likely irritate the user. Improvements of the synthetic face are already in progress, as discussed in chapter 2.2. The features under further development are:

- The speech recognizer (which must be faster)
- The basic parameters controlling the animation (in order to improve the quality of the movements)
- The development of emotional expressions and wrinkles (to achieve movements that are more natural)

Another important aspect to consider is the price of the equipment (computer, modem, speech recognizer). If the product is too expensive, no user will appreciate it.

In the forthcoming evolution of SYNFACE, a few aspects must be considered. Is the best solution the introduction of latency of the total signal, thereby risking problems with turn taking? Or is an immediate audio signal more desirable, causing asynchrony between the two modes? To be able to answer these questions, further testing must be conducted, preferably on hearing impaired subjects. If the asynchrony case is chosen, an optimal delay value must be found: fast working hardware and well functioning software is necessary to diminish the delay as much as possible. The results of the present study suggest a maximal delay around 50 ms. However, there is a possibility that the tolerance is greater than shown here. Therefore, an investigation of delays in the interval [-50, -175] would be interesting to perform.

***

6 ACKNOWLEDGEMENTS

Special thanks to,

Björn Granström - *who supervised without commanding, thus letting me plan, perform and complete the project independently.*
Jonas Beskow - *who assisted me with the software I needed for the experiments.*
Giampiero Salvi - *for nice company in the office (now you will finally get rid of me...)*
Mattias Heldner - *for the patience when helping me with the statistics.*
Inger Karlsson - *for assisting me in various issues during the project.*
Robert Mannell - *for improving the report linguistically.*
Ve - *for many mails and faxes in short time.*
Per, Boris, Gunilla, Mikael, Erik, Alec & Peter - *for nice lunches, breaks and discussions.*
The subjects - *who entered for the experiments with short delay and with a happy attitude.*
The SYNFACE group and all the people at CTT - *for a warm and friendly atmosphere.*

My dear family & sweet friends - *for always being there for me.*

7 REFERENCES

Agelfors E, Beskow J, Dahlquist M, Granström B, Lundeberg M, Spens KE, Öhman T (1998). Synthetic faces as a lipreading support. In *Proceedings of the International Conference on Spoken Language Processing, Sydney, Australia, 7: 3047-50.*

Arlinger S, Brorsson B, Lagerbring C, Leijon A, Rosenhall U, Schertstén T (2003). Hörselproblem hos vuxna – nytta och kostnader för hörapparater. SBU: Statens Beredning för medicinsk Utvärdering.

Bernstein LE, Benoit C (1996). For speech perception by humans or machines, three senses are better than one. *ICSLP 96 Fourth International Conference on Spoken Language Processsing,* 3: 1477-1480, Philadelphia, PA, 1996.

Beskow J (to appear). Talking Heads - Models and Applications for Multimodal Speech Synthesis. Doctoral Dissertation (to appear in June 2003), KTH, Stockholm, Sweden.

Campbell R, Dodd B (1980). Hearing by Eye. *Quarterly Journal of Experimental Psychology* 32: 85-99.

Dahlquist M (2002). Utveckling av meningsmaterial för testning av personer med cochleaimplantat. Rapport till Tysta Skolan.

Dixon NF and Spitz L (1980). The detection of auditory visual desynchrony. *Perception,* 9: 719-21.

Engwall O (2002). Tongue Talking – Studies in Intraoral speech synthesis. Doctoral Thesis. Department of Speech, Music and Hearing, KTH, Stockholm.

Fisher CG (1968). Confusions among visually perceived consonants. *Journal of Speech and Hearing Research*, 11, 796-804

Grant KW, Greenberg S (2001). Speech intelligibility derived from asynchronous processing of auditory-visual information. *Proc. AVSP 2001 International Conference on Auditory-Visual Speech Processing*, Scheelsminde, Denmark, 132-137.

Isaacs E, Tang JC (1993). What video can and can’t do for collaboration: a case study. In *Proceedings of the ACM Multimedia Conference,* Anaheim, CA.

Le Goff B, Guiard-Marigny T, Cohen MM, Benoît C (1994). Real-time analysis-synthesis and intelligibility of talking faces. *Proceedings of the second SCA/IEEE Workshop on Speech Synthesis*, New Paltz, New York, USA.

Lidestam B, Lyxell B, Lundeberg M (2001). Speechreading of synthetic and natural face. *Scandinavian Audiology* 30: 89-94.

MacLeod A, Summerfield Q (1990). A procedure for measuring auditory and audio-visual speech-reception thresholds for sentences in noise: rationale, evaluation, and recommendations for use. *British Journal of Audiology* 24: 29-43.

Massaro DW (ed.) (1980). Perceptual events and time. In: *Perceiving talking faces, from speech perception to a behavioral principle.* London: The MIT Press*,* 72-89.

Massaro DW, Cohen MM (1993). Perceiving asynchronous bimodal speech in consonant-vowel and vowel syllables. *Speech Communication* 13: 127-134.

McGrath M, Summerfield Q. (1985) Intermodal timing relatings and audio-visual Speech Recognition by normal-hearing adults. *Journal of the Acoustic Society of America* 2: 678-685.

Oerlemans M, Blamey P (1998). Touch and auditory-visual speech perception. In: Campbell R, Dodd B & Burnham D, eds. *Hearing by Eye II*. East Sussex: Psychology Press Ltd, 267-281.

Pandey PC, Kunov H, Abel SM (1986). Disruptive effects of auditory signal delay on speech perception with lipreading. *The Journal of Auditory Research* 26: 27-41.

Pihl E (2003). Bottlenecks in the Synface telephone. Master thesis at CTT.

Ruhleder K, Jordan B (2001). Co-constructing non-mutual realities: delay-generated trouble in distributed interaction. *The International Journal of Computer Supported Cooperative Work* 10(1):113-138.

Schomaker L, Nijtmans J, Camurri A, Lavagetto F, Morasso P, Benoît C, Guiard-Marigny B, Le Goff B, Robert-Ribes J, Adjoudani A, Defée I, Münch S, Hartung K, Blauert J (1995). A taxonomy of multimodal interaction in the human information processing system. A report of the ESPRIT PROJECT 8579 MIAMI, page 38. Also to be found at http://hwr.nici.kun.nl/~miami/taxonomy/taxonomy.html

Sellen AJ (1992). Speech patterns in video-mediated conversations. *Proceedings of CHI '92*, Monterey, CA, 49-59.

Siciliano C, Williams G, Faulkner A (2002a). SYNFACE Deliverable D3-1: Performance of Synthetic face driven by annotated speech.

Siciliano C, Williams G, Beskow J, Faulkner A (2002b). Evaluation of a synthetic talking face as a communication aid for the hearing impaired. *Speech, Hearing and Language: Work in Progress* 14: 51-61.

Smeele PMT (ed.) (1994). *Perceiving Speech: Integrating Auditory and Visual Speech.*

Tang JC, Isaacs E (1993). Why do users like video? Computer Supported Cooperative Work. *Journal of Collaborative Computing*, 1(3): 163–196.

Ward K (2002). Användargränssnitt för Synface-projektet. Master thesis at CTT.

8 APPENDIX

APPENDIX 1 – Information to subjects in Swedish
APPENDIX 2 – Tables from the statistical analysis (2 pages)
APPENDIX 3 – Individual charts (4 pages)

APPENDIX 1 – Information to subjects in Swedish

”EXPERIMENT WITH ASYNCHRONY IN MULTIMODAL SPEECH COMMUNICATION”

Försöket ingår i ett projekt som heter SYNFACE, vars syfte är att utveckla ett hjälpmedel för hörselskadade vid telefonkommunikation. Den inkommande talsignalen styr läpp- och tungrörelserna i ett syntetiskt ansikte vilket möjliggör läppavläsning och därmed underlättar kommunikationen för den hörselskadade. I projektets nuvarande prototyp uppstår dock en tidsfördröjning mellan talsignalen och bilden, vilket kan förvirra användaren. Ändamålet med denna studie är att undersöka när tidsfördröjningen blir så stor att nyttan med det syntetiska ansiktet uteblir.

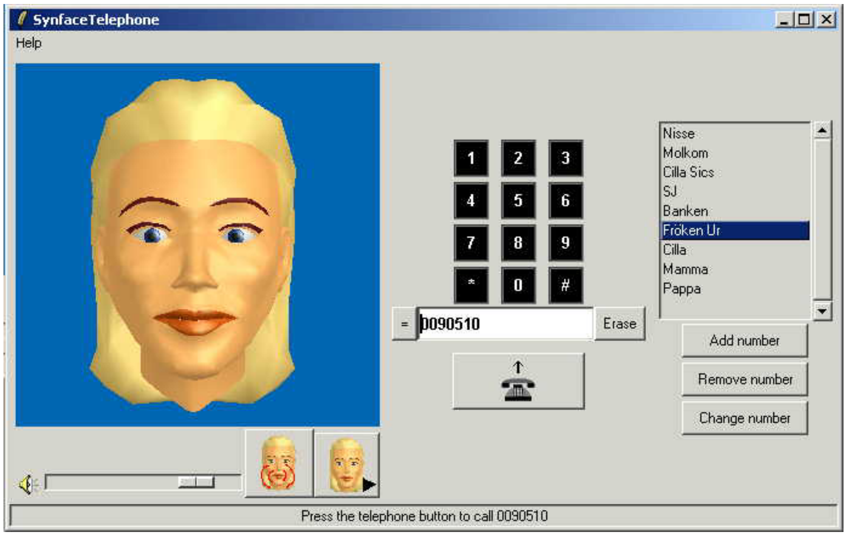
TILLVÄGAGÅNGSSÄTT

Du kommer att se ett ansikte och höra en röst. Dessa två stimuli är dock inte synkroniserade i tid för alla testvillkor, utan kan ha olika tidsfördröjningar. Ibland kommer ljudet före bilden, ibland efter. Ansiktet på skärmen kan vara antingen naturligt eller syntetiskt och ljudet är behandlat med en störning för att simulera en hörselskada. Varje kombination av ansikte och tidsfördröjning utgör ett testvillkor. Sju av villkoren är med naturligt ansikte och fyra med syntetiskt. Dessutom består ett testvillkor av endast ljud, då visas en bild på en högtalare istället för ansiktet.

Materialet består av 14 korta meningar för varje testvillkor. Villkoren spelas upp i slumpmässig ordning och är uppdelade på två lika stora block, d v s med 6 villkor i varje. Efter varje stimulus säger du högt vad du uppfattade och försöksledaren antecknar ditt svar. Innan testet påbörjas, genomförs en övningsomgång för att du ska veta hur allt fungerar. I denna får du ett smakprov på varje testvillkor, något som kan vara förvirrande på grund av de många varianterna av stimulus. Detta blir dock inget problem då det riktiga testet körs, eftersom varje testvillkor genomförs i omgångar om 14 meningar.

ATT TÄNKA PÅ

Säg exakt vad du uppfattat, även om det inte finns något logiskt sammanhang mellan orden. Om du bara uppfattar ett ord – säg det!



0090510

Press the telephone button to call 0090510

Synface Telephone

/TMH, Marie Molander, sep -02

I. ANOVA, investigating the delay effect

	DF	Sum of Squares	Mean Square	F-Value	P-Value	Lambda	Power
Delay	6	16605.489	2767.582	5.019	.0002	30.117	.993
Residual	77	42455.633	551.372				

II. Fisher's PLSD post hoc test

	Mean Diff.	Crit. Diff.	P-Value	
D-300, D-175	-1.389	19.089	.8852	
D-300, D-50	-24.769	19.089	.0117	S
D-300, D0	-29.861	19.089	.0026	S
D-300, D50	-37.269	19.089	.0002	S
D-300, D175	-30.324	19.089	.0022	S
D-300, D300	-9.028	19.089	.3493	
D-175, D-50	-23.380	19.089	.0170	S
D-175, D0	-28.472	19.089	.0040	S
D-175, D50	-35.880	19.089	.0003	S
D-175, D175	-28.935	19.089	.0034	S
D-175, D300	-7.639	19.089	.4280	
D-50, D0	-5.093	19.089	.5968	
D-50, D50	-12.500	19.089	.1961	
D-50, D175	-5.556	19.089	.5639	
D-50, D300	15.741	19.089	.1047	
D0, D50	-7.407	19.089	.4421	
D0, D175	-.463	19.089	.9616	
D0, D300	20.833	19.089	.0328	S
D50, D175	6.944	19.089	.4710	
D50, D300	28.241	19.089	.0043	S
D175, D300	21.296	19.089	.0293	S

III. ANOVA, investigating the effect of leading channel

	DF	Sum of Squares	Mean Square	F-Value	P-Value	Lambda	Power
Lead/lag	1	5093.021	5093.021	7.642	.0073	7.642	.789
Residual	70	46650.806	666.440				

IV. ANOVA, investigating the effect of type of face

	DF	Sum of Squares	Mean Square	F-Value	P-Value	Lambda	Power
Face	1	3.938	3.938	.005	.9411	.005	.051
Residual	94	67564.783	718.774				

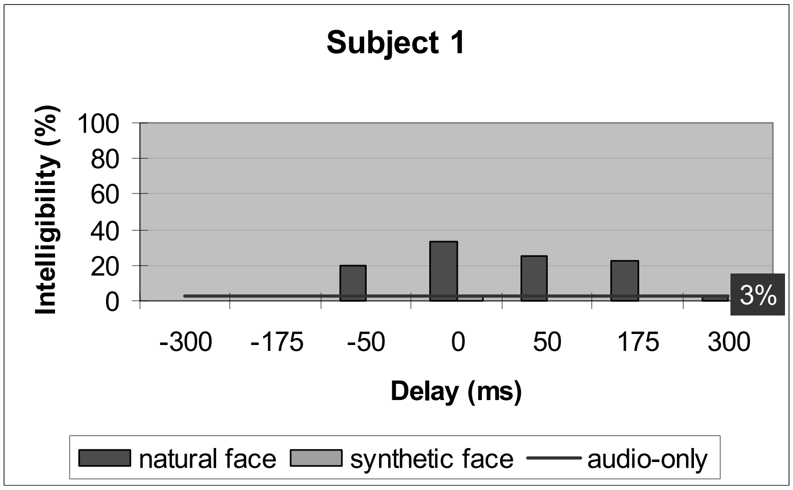
V. ANOVA, investigating the effect of sequence

	DF	Sum of Squares	Mean Square	F-Value	P-Value	Lambda	Power
Sequence	11	9312.629	846.603	1.166	.3166	12.828	.613
Residual	132	95824.331	725.942				

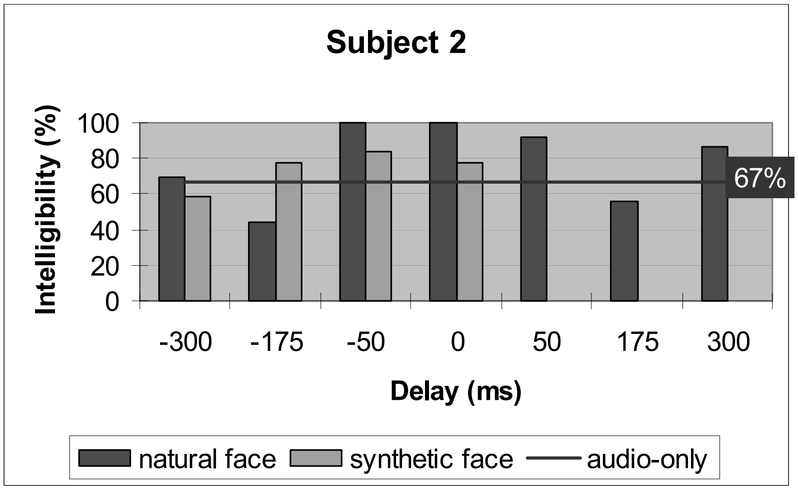
VI. ANOVA, investigating the effect of sentence list

	DF	Sum of Squares	Mean Square	F-Value	P-Value	Lambda	Power
List	11	4542.824	412.984	.542	.8716	5.961	.282
Residual	132	100594.136	762.077				

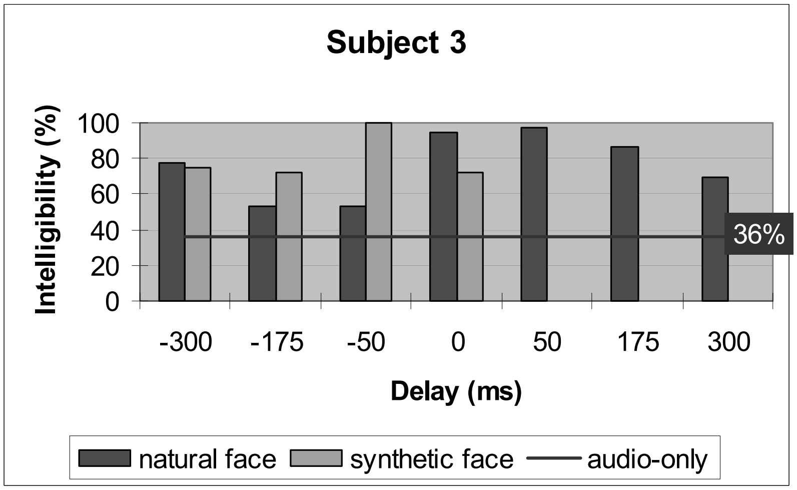
APPENDIX 3 (page 1) – Individual charts



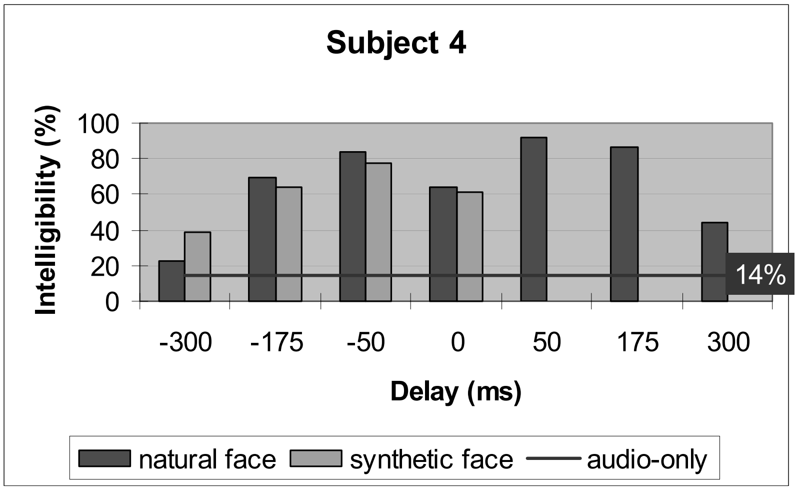
Sequence	Intell(%)
synt 0	1
nat 0	33
nat 175	22
nat -300	0
nat 300	3
synt -300	0
nat 50	25
nat -175	0
audio-only	3
synt -50	0
synt -175	0
nat -50	19



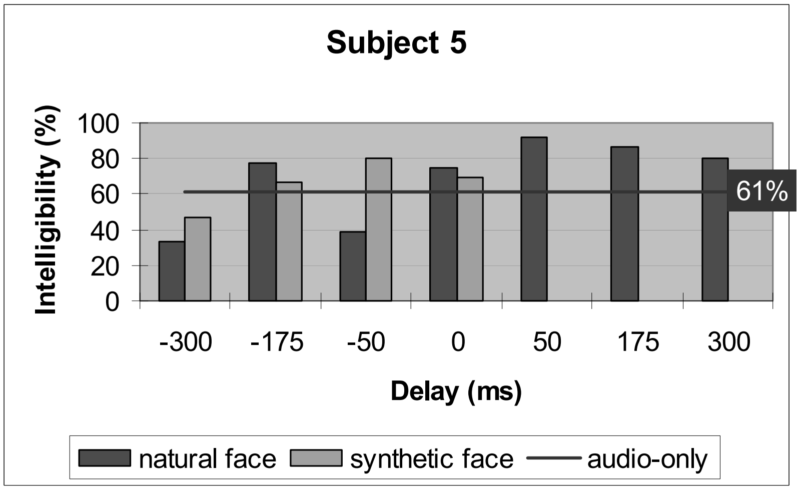
Sequence	Intell(%)
nat 175	56
nat 0	100
nat -175	44
nat -50	100
synt -300	58
synt 0	78
nat 50	92
synt -50	83
nat -300	69
nat 300	86
synt -175	78
audio-only	67



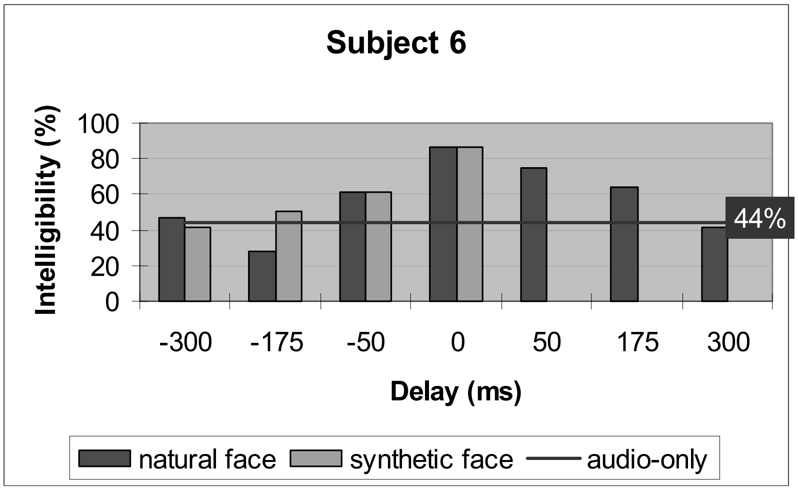
Sequence	Intell(%)
nat -50	53
synt 0	72
nat 0	94
nat -175	53
nat 50	97
nat 300	69
audio-only	36
synt -175	72
synt -50	100
nat -300	78
synt -300	75
nat 175	86



Sequence	Intell(%)
audio-only	14
nat -300	22
synt 0	61
synt -175	64
nat 300	44
synt -50	78
nat 0	64
nat -50	83
nat 175	86
nat -175	69
synt -300	39
nat 50	92

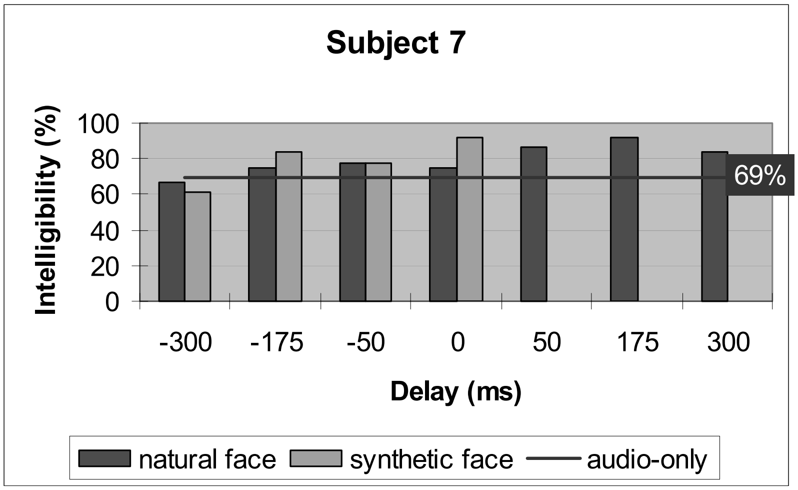


Sequence	Intell(%)
nat -50	39
synt 0	69
synt -300	47
nat 175	86
synt -175	67
synt -50	81
nat -300	33
nat 0	75
nat -175	78
nat 50	92
nat 300	81
audio-only	61

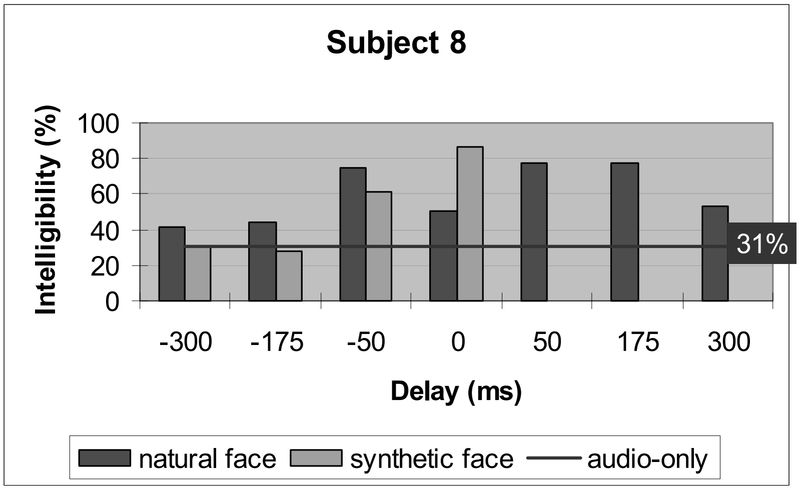


Sequence	Intell(%)
nat -50	61
nat 175	64
nat 300	42
nat -175	28
audio-only	44
synt 0	86
nat 50	75
synt -300	42
nat 0	86
nat -300	47
synt -175	50
synt -50	61

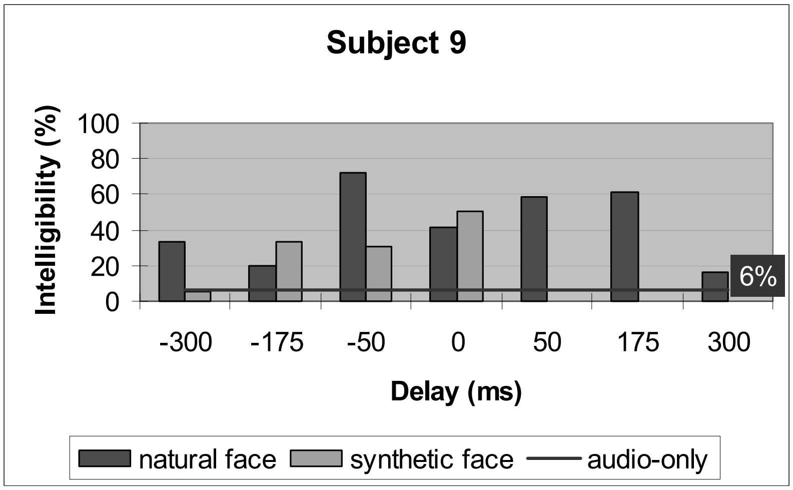
APPENDIX 3 (page 3) – Individual charts



Sequence	Intell(%)
nat 0	75
synt -300	61
synt 0	92
nat -50	78
nat 50	86
audio-only	69
nat -300	67
nat -175	75
nat 175	92
nat 300	83
synt -50	78
synt -175	83

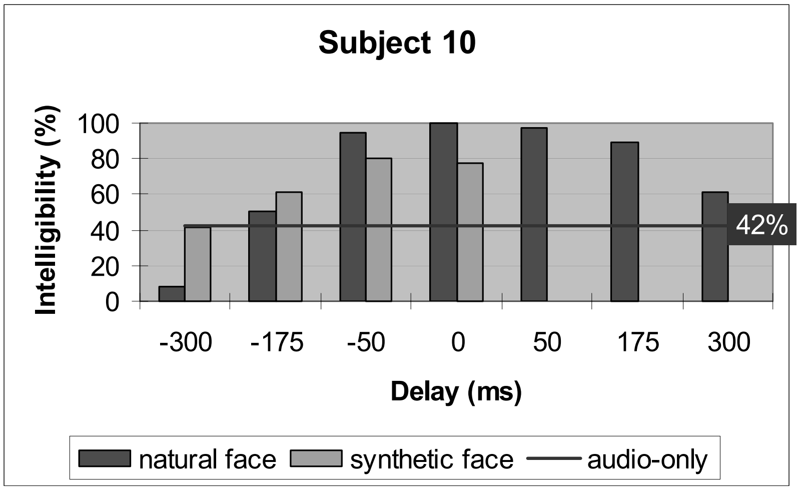


Sequence	Intell(%)
nat 0	50
nat 50	78
synt -300	31
synt -50	61
nat -300	42
synt -175	28
nat -175	44
nat 175	78
nat -50	75
audio-only	31
nat 300	53
synt 0	86

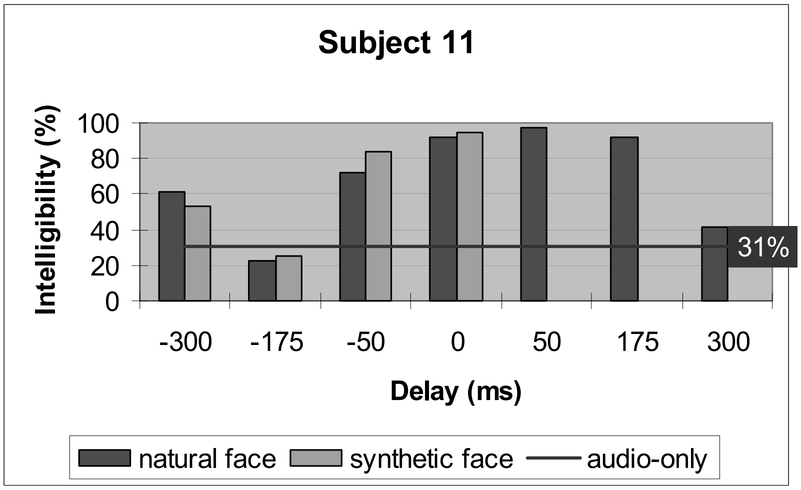


Sequence	Intell(%)
nat 300	17
synt -300	6
audio-only	6
nat 50	58
nat 0	42
nat -50	72
synt -50	31
nat -175	19
nat 175	61
synt -175	33
synt 0	50
nat -300	33

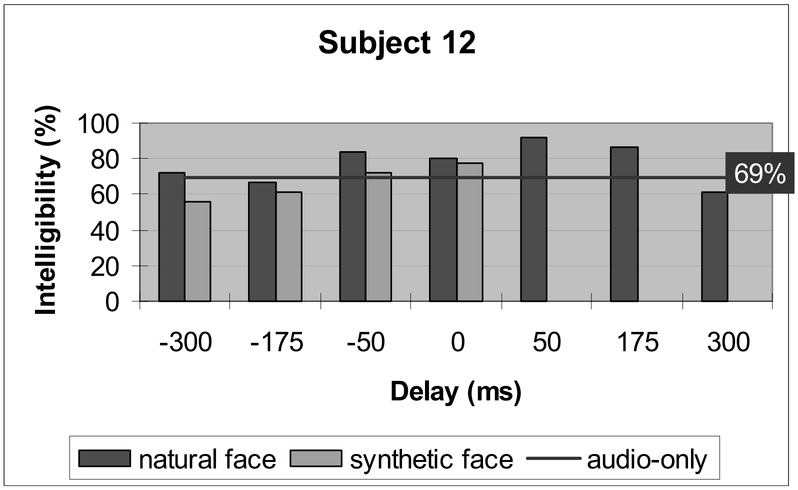
APPENDIX 3 (page 4) – Individual charts



Sequence	Intell(%)
nat -300	8
synt -175	61
nat -50	94
synt -300	42
nat 50	97
nat 175	89
nat 300	61
audio-only	42
nat 0	100
nat -175	50
synt 0	78
synt -50	81



Sequence	Intell(%)
audio-only	31
nat -175	22
synt -50	83
synt -175	25
nat 175	92
nat -50	72
nat 50	97
synt 0	94
nat 300	42
synt -300	53
nat -300	61
nat 0	92



Sequence	Intell(%)
synt -175	61
nat 50	92
nat 300	61
nat 0	81
synt -50	72
audio-only	69
nat 175	86
nat -300	72
nat -175	67
synt -300	56
synt 0	78
nat -50	83