Acceptance of Synthetic speech in Multi-modal Information Retrieval

A Master Thesis

By Andrew Glenn Moores

Department of Information Science

University of Bergen

October 2009

PREFACE

This master thesis is an end to my masters' degree at the University of Bergen, and is written within the field of Information Science.

Looking at synthetic speech from a user's point of view and knowing that there is room for improvements, has been a very interesting approach. Although the work with the thesis has been time-demanding and complicated, I feel that I am left with a greater understanding of a research topic that I had limited knowledge of before I started. The fact that my results and conclusions might contribute to an ongoing field of research has given me motivation throughout the whole project.

The biggest challenge I faced in this process was the lack of similar data that I could compare my own gathered data with. Because of this, it was difficult to draw any concrete conclusions as to what could be improved with synthetic speech. However, the participants in my research have been of great help and without them I wouldn't be able to draw any kind of conclusions at all. I would therefore like to thank every participant for their time and valuable feedback.

Last, but not least, I want to thank my guidance counselor at UiB, Associate Professor Joan C. Nordbotten. She has contributed with many good ideas and helpful comments throughout the entire process of this thesis.

Bergen, October 21st, 2009

Andrew Glenn Moores

Table of Contents

SUMMARY

This thesis concentrates on the use of synthetic speech, utilizing a TTS engine, by presenting users with an audio description of historical monuments. The main focus of the thesis is to find out in what way users think synthetic speech can be improved so to increase their acceptance of synthetic speech in future uses.

The results gathered from participants during testing show indications of several factors leading to better acceptance. These factors involved more exposure to synthetic speech, slower speech speed from the “speaker”, less choppiness (more naturalness) and better pronunciation with regards to voice pitch depending on word context, and recognition of punctuation.

Based on these results, there seems to be clear indications of this thesis’ proposed hypothesis being false. It was discovered that solely changing the *duration* of audio clips did not play a major role in improving user acceptance. Rather, results indicated a higher acceptance from users based on more *exposure* to the presented synthetic speech. In other words, for the majority of participants, the more they listened to the synthetic speech, the more positive their responses became. In addition to exposure, factors based on quality improvements were also recommended as other ways of increasing acceptance.

1 INTRODUCTION

Imagine being a tourist in a country full of historical landmarks such as building, statues, markets and so on. Being a tourist, it is almost certain that it is the first time they have ever seen the landmarks and almost equally certain that they do not know all the historical facts about them. Because of this, a tourist might want to find out more details about a landmark after being at the site. However, searching for information without being familiar with the background of the landmark, and therefore not knowing exactly what to search for, could prove rather difficult.

It is also possible to receive information on a landmark, or similar landmarks, while one is still at the site looking at it. This would be more practical for the tourist in the sense that the information becomes available there and then, and not later on. A mobile phone can be used for this purpose, so that one can search the Internet without needing a computer. However, the above mentioned problem still exists. In the case of one not knowing the name or background of a landmark, it is hard to know what to search for. The tourist is left only with the choice of describing the landmark with different keywords while trying to find information on it. A better method would be to use an actual picture of the landmark as a search term. This will make it easier to identify the landmark and finding information describing and explaining it.

The type, or format, in which information is given back to a tourist may vary in desire from user to user, be it in the form of text, audio, images, or a combination of these. One matter that arises is that of convenience. Is it more convenient to walk around looking at landmarks and at the same time be looking at one's mobile devise *reading* facts about the landmarks? Or would it be more convenient to look at landmarks and at the same time be *listening* to someone else talking about the facts and thus not needing to look away?

With today's advancements in mobile technology, accessing information has become much easier. Instead of having to sit in front of a computer one can simply use a mobile phone, or any other handheld device, that can connect to the Internet and search for information on the spot. Along with the possibility to search for information no matter where one might be (as long as it is possible to connect to the Internet), it is also a great advantage that these same

devices are able to accept information in just about any data form, be it a video clip, audio clip, picture or text. There are of course some drawbacks though to mobile devices. In the case of a tourist, searching for information on a mobile phone may be easy, however, factors such as the size of the mobiles' screen or the possible glare from too much light outside may cause problems. A known case among elder people and those with poor vision is the possible cause of stress and/or irritation with not being able to see what is on small mobile telephone screens. As for possible glares from bright light outdoors, this is a problem any user can encounter, but isn't the worst of situations to get out of.

It should be mentioned that tourists, such as in this case, can be both local (sightseeing in their home country) and foreign (sightseeing in a different country than their home country). Although some foreign tourists are willing to use their mobile devices, such as mobile phones, while abroad, it is more likely that local tourists to a larger extent are willing to use their mobile phones to try out such a system. This is mainly due to the fact that local tourists pay a lower price for Internet services than foreign tourists[1]. A more detailed explanation of possible users is presented in chapter 3, section 3.1.

The rest of this chapter will be presented in the following order: section 1.1 introduces possible problems to recorded texts and introduces text-to-speech technology as a possible solution. Section 1.2 gives a brief overview of the main focus area for this thesis, presenting the proposed research question and hypothesis. Finally, section 1.3 gives a detailed overview of how data collection, analysis and evaluation are planned on being conducted.

1.1 ISSUES AROUND TEXT-TO-SPEECH (TTS) CONVERSION

As mentioned earlier it may be desired by users, for example tourists as in the previous section, to listen to someone talking about relevant information in a certain situation. In other words, instead of reading the information about a landmark themselves, they could have someone else read it for them.

[1] These prices are constantly changing, so the situation might be different in the future.

One solution would be to have a guide at every landmark talking to tourists, although this is highly unlikely to happen. An alternative would be to have a recording with facts and information on a landmark available for anyone wanting it. This alternative is used in many situations, such as sightseeing buses and in some museums[2]. A problem with the bus example is that they're only able to listen to the recording in a certain spot before driving along to the next landmark. It is not possible to move around at the site and look at the landmark from different angles. As for the museum example, the availability of the recording may be considerably low in the case of a long line-up of people wanting to listen. Another, and perhaps better, alternative would be to download *audio clips*, with recorded information about a landmark, on the spot, onto a mobile device such as a mobile phone. Such recordings could be saved in a database available for anyone at any time. If a tourist would prefer to explore the landmarks by foot instead of by bus, he or she will then have access to the relevant information all the time. Concerning the museum example, it wouldn't be necessary to get in a line-up because all the information would be saved on the mobile phone.

There are (at least) two ways for audio clips to be generated for such a cause. One of them involves manually recording a text, and the other one involves an automatic conversion. It's assumed that there are not very many ways of manually recording texts. Depending on the context, a person may be hired for reading and recording texts such as audio books, or in the case of tourists, perhaps audio descriptions of historical monuments. Although the quality of the audio recording could be very good, there are other issues that make the automatic conversion more appealing. If a company focusing on, say, producing recordings of historical monuments primarily based on the English language, suddenly wanted to expand and translate their English recordings to more languages, what should they do? One solution could be to have the original recorder learn the new languages they wish to now have recordings in. This may seem a bit drastic and highly unlikely. A different alternative could be to hire new recorders that are fluent in these languages. This alternative seems much more likely to be done, however, then the issue of cost appears. The more people needed for recording texts, the higher the costs will be for that company as they would have to start paying more and more people for the job. A third issue that arises is time. Should a company need to produce new recordings in short time, they will be dependent on always having someone that can

[2] Audio Description's homepage.

record texts in certain languages. Should someone quit or fall ill, this time demand may cause problems.

Because of these issues, automatic conversion stands out as a better alternative for generating audio clips. There exist many tools for automatically converting texts to speech, or audios. Some examples of these tools are “NaturalReader”, “Sayvoice”, “ReadPlease”, and “VoiceMX STUDIO”[3]. These tools are also known as text-to-speech engines, or TTS, defined and explained in greater detail in chapter 2. How the audios are made can vary from tool to tool. The most common method is for the tool to read the text, compare the words to a database or collection of recorded words, and concatenate, or sting together, these to create an audio version of the text[4]. Compared to manually recording texts, TTS engines may be both more time- and cost efficient. It would no longer be necessary to have several employees for recording different languages, but rather just invest in multi-language TTS engines. Although it may cost to implement such a tool, it would pay off in the long run. As for time aspect, the TTS engine could be used at any time, eliminating the problem of having to wait for someone to be able to record a text. In addition to this, the actual time needed for recording text would be significantly less as the TTS engine would perform the conversion of text to speech much quicker than a human.

1.2 RESEARCH PROJECT

Regarding the field of TTS in this thesis, the focus will be on the use of an existing, functioning, tool rather than the architecture around how the process is executed. The reason for this is that users do not have to execute the actual conversion of text to audio themselves. It is therefore not necessary to explain in depth how the conversion will be conducted. In other words, it’s the *product* of TTS conversion that is under investigation, not the *architectural* side of it. Issues around improvements to the quality of the produced speech will be presented in later chapters (4 and 5).

This thesis focuses on the following proposed research question:

[3] Respectively NaturalSoft’s homepage, Sayvoice’s homepage, ReadPleases homepage and Tanseon System’s homepage. Actual links are listed in the bibliography under “Internet Sources”.
[4] A good overview of methods used for developing TTS engines can be found in “Review of text-to-speech conversion for English” (Klatt, 1987), where many methods for how conversion is executed are brought to light.

RQ: "How can the acceptance of synthetic speech be improved?"

The research question is based on the assumption that a recorded *human voice* would be preferred over *synthetic speech*. This assumption was confirmed by a minor test session, where a sample of a human recording of a text was played to 10 people followed by playing the same text with the use of a synthetic voice. The reactions given were that the human recording was preferred. This was anticipated from the start, but the testing was performed to document this assumption. Knowing that the participants here did *not* prefer synthetic speech made it possible to refine the research question, so that focus was solely put on the aspect of how synthetic speech could be *improved*.

The following hypothesis was formed as a basis for exploring the research question:

HYP: "Shorter audio clips make synthesized speech easier to listen to and thereby more accepted."

"Shorter segmented audio clips" simply refers to the length, or duration, of the converted texts. As will be explained in greater detail in chapter 4, the audio clips that will be presented during testing are divided into three categories; short, medium, and long.

Should this hypothesis be proven true it may be possible to conclude two important issues. Firstly, the *quality* of synthetic speech may be the prime reason for the lack of acceptance by users. When listening to a *long* audio clip and thereby being exposed to synthetic speech of poor quality for a longer period of time, users may "fall off track" or lose interest in listening to the clip. Secondly, following the first issue, by offering *short* clips of a converted text, it may be easier for users to listen to a whole description without growing tired and not wanting to listen to a whole clip. The reason for this is that they can choose whether to continue listening or to stop listening, or at least take a short break between each (short) clip.

1.3 RESEARCH APPROACH

As this thesis focuses on how users experience listening to audio clips of synthetic speech, there are many ways in how research can be conducted. Figure 1.1 shows an overview of such methods from Kjeldskov & Graham (2003), with brief information on strengths, weaknesses and how the methods can be used.

	Method	Strengths	Weaknesses	Use
Natural setting	Case studies	Natural settings Rich data	Time demanding Limited generalizability	Descriptions, explanations, developing hypothesis
	Field studies	Natural Settings Replicable	Difficult data collection Unknown sample bias	Studying current practice Evaluating new practices
	Action research	First hand experience Applying theory to practice	Ethics, bias, time Unknown generalizability	Generate hypothesis/theory Testing theories/hypothesis
Artificial setting	Laboratory experiments	Control of variables Replicable	Limited realism Unknown generalizability	Controlled experiments Theory/product testing
Environ-ment independ-ent setting	Survey research	Easy, low cost Can reduce sample bias	Context insensitive No variable manipulation	Collecting descriptive data from large samples
	Applied research	The goal is a product which may be evaluated	May need further design to make product general	Product development, testing hypothesis/concepts
	Basic research	No restrictions on solutions Solve new problems	Costly, time demanding May produce no solution	Theory building
	Normative writings	Insight into firsthand experience	Opinions may influence outcome	Descriptions of practice, building frameworks

Figure 1.1 – Summary of research methods (Extracted from Wynekoop & Conger [11]) (Kjeldskov & Graham, 2003))

Out of the methods mentioned in the figure above, the most relevant ones for this thesis are field studies, laboratory experiments, and survey research. This is based on several factors. First, the use of a field study and laboratory experiment are used to examine whether or not the environment surrounding participants during testing reveal any differences in results gathered. Secondly, survey research can provide valuable feedback from participants giving very descriptive data in relation to the proposed research question and hypothesis.

1.3.1 DATA COLLECTION

Data will be collected through a field study and a laboratory experiment using a survey/questionnaire in both[5]. In other words, there will be two sessions of testing; respectively a real-scenario test and an in-house test.

[5] A definition of "field studies" and "survey research" can be found in Kjeldskov & Graham (2003).

Testing will be conducted through a series of individual meetings with the testers. It is estimated that somewhere around 10 to 15 male and female participants will be used throughout all the testing, adding up to a total of 20 to 30 testers. The participants are all 1st year University students. The primary goal of choosing 1st year students as testers is to simulate possible users, as explained in section 3.2. However, another goal is to insure that the testers have relatively little knowledge of the different alternatives available within TTS products. Because the testers will have limited knowledge of possible better alternatives than the chosen TTS tool used in this thesis, one avoids the fact that their knowledge can affect their responses. Hence, the quality of their feedback will increase.

Once a set of stored texts has been converted into audio clips (the chosen TTS tool will be presented in chapter 3, section 3.1) a pre-selected set of the automatically generated audio clips will be presented to each participant for testing. The only criterion for each set was that there should be exactly 2 audio clips from each of the 3 duration categories; short, medium and long. More detail into the construction of these sets is presented in section 3.2.2 After the users have listened to the audios they will then be asked to respond to a questionnaire about different quality marks on the audio clips. The questionnaire consisted of the following questions[6]:

Audio Quality Survey:

1) Was the speech smooth or choppy?
2) Was the speech difficult to understand (was it clear/unclear what was being said)?
3) Was this form of speech tiring to listen to?
4) Could you listen to this form of speech for a longer period of time than this without wanting to stop?
5) How do you think this form of speech could be improved?
6) Any other comments: _____

During the process of collecting data, participants, depending on which test session they are a part of, will be asked to either go to the heart of Bergen, Fisketorget (the real-scenario test), or to sit in front of a computer (the in-house test). The participants that go to Fisketorget will

[6] The questionnaire used during testing may also be found at the end of this thesis in appendix C.

get a mobile phone with pre-selected clips on it, while the other participants will get the pre-selected clips on the computer. Further, in both test sessions they will be given headphones to listen to the audio clips with. The reason for this choice, during the real-scenario test session, is based on the assumption that using headphones will increase the chance of hearing all of what is said in the audios – seeing as standing in the middle of a city is not exactly noise free. In other words, using headphones may block out many distracting sounds such as cars, animals, or other people walking by and talking, yelling and so on. During the in-house test session the headphones are also supposed to block out any surrounding noise, but in addition they are used to give the participant a feeling of how the audio clips would be used in a real situation.

As mentioned, the testers will be given pre-selected audio clips to test. These clips, each consisting of an independent topic (churches, buildings, statues, and so on), will be presented in different durations, to examine how the users will react when presented with clips of varying duration. The varieties of the audio clips will be in the form of:

- “Short” clips, between 6-25 seconds
- “Medium” clips, between 26 and 45 seconds and
- “Long” clips, over 45 seconds

As for the motive of how the various boundaries between short, medium and long clips was derived, chapter 3, section 3.1 will go into greater detail on this.

Testers will be divided into groups and then test each of the variants, however, each group will test the sample variants in different orders. This is done to examine if testers react to the same descriptions in different ways depending on what *order* they are listened to. A framework for how this will be executed is as followed:

- First male and female participants: long, medium, short of clip set 1
- Second male and female participants: short, medium, long of clip set 1
- Third male and female participants: long, medium, short of clip set 2

- And so on until there are no more participants.

This framework goes for both test sessions that will be used.

Once the testing is completed, the collected data will be analyzed and evaluated in an attempt to prove the validity of the proposed hypothesis.

1.3.2 PROCESSING THE DATA

Once all the data is collected from testing, the following steps will be conducted for analysis and evaluation.

First and foremost, the results will be organized into two separate groups; female and male results. Although the gender of the participants plays no immediate role in the research question and hypothesis posed in this thesis, this split was made simply to investigate whether or not there was an actual difference in responses from each gender. From here the results will be further divided into more groups where each group represents a question from the questionnaire. In other words, all data collected from female participants on question 1 will be presented first, followed by female results from questions 2, 3, and so on. When the female results have been presented, the male results will be presented in the same manner; splitting results from each question into separate groups.

The next step will be to actually analyze and evaluate the feedback from each question to get an understanding of what users actually think about the use of synthetic speech. There are two main goals in this process. Firstly, to find out whether or not participants experience any differences in acceptance towards synthetic speech depending on the various *durations* of the clips. And secondly, to examine how the use of computerized voices can be *improved* in order to increase the acceptance of such speech among users.

During the analysis of results, summary data will be presented in the form of values extracted from the questionnaire ranging from 1 to 7, depending on the number of options given for each question. Averages will then be calculated to try and compare the results from female and male participants to see if there are any major differences in results with regards to

gender. A discussion on the age and educational level of participants will also be presented for similar reasons. Finally, conclusions will be devised in an attempt to answer whether or not the proposed hypothesis is valid.

1.4 CHAPTER SUMMARY

In chapter 1 an introduction has been given as to how information can be presented to users, in this case tourists. Following the introductory case, background information for this thesis was presented. In particular, it was focused on issues around TTS conversion and some of the benefits of using such tools compared to manually recording texts. The main benefit was the fact that TTS engines can be both more time- and cost efficient. The second area of this chapter presented the actual proposed research question and its' corresponding hypothesis, respectively the acceptance of synthetic speech and whether or not the duration of clips plays a role in this. Following this was the presentation of the research methods used in attempting to validate/falsify the proposed hypothesis. Data collection is planned to be conducted by utilizing an in-house test and a real-scenario test. Audio clips of short, medium and long durations are to be presented to participants, followed by a questionnaire to be answered after listening to two clips of each duration category. As for the processing- and evaluation plan, this is going to be conducted by first splitting all results into two groups; female and male, and then splitting the results into groups consisting of the answers from each question posed in the questionnaire. During the analysis of the results, trends and indications will be gathered and presented, along with summary data of the responses of participants.

The rest of this thesis is presented in the following way. Chapter 2 will give an overview of the *theory* surrounding the main aspects of this thesis. Following this, chapter 3 goes into more detail on the *data collection framework*. In chapter 4, the results from testing will be *analyzed* and possible indications and trends of the participants' responses will be presented. Finally, in chapter 5, *evaluation* of the testing results and proposed hypothesis will be produced, followed by *conclusions* and possible topics of interest to look into for *future research*.

2 THEORY

The research area of this thesis is based within the field of information retrieval and in the field of TTS synthesis. The following chapter will present the theoretical framework that is considered relevant in regards to these fields. Although all the theory and definitions presented here might not be used in a direct manner, it still seems necessary to include it in order to give a proper and complete overview of the different topics.

Starting off, section 2.1 takes a closer look into the different forms of data that exists, while section 2.2 gives an overview of different ways that information can be represented followed by some of the key concepts within information retrieval, and some of the different ways of performing searches on various forms of data with various forms of data. Finally, section 2.3 defines key concepts within TTS, and goes into detail as to how TTS has previously been evaluated and used.

2.1 FORMS OF DATA

Information can be derived from a variety of different forms of data. Several forms of data exist, however those that are most relevant for this thesis are text, images and audio. Although text can be defined as sets of specifically arranged symbols, or characters, an alternative explanation is that text is a human- and/or computer-readable sequence of characters creating words, sentences, or paragraphs produced and stored on a medium. Text can be the contents of a book, codes and numbers on a purchase receipt, and many other things. In relation to this thesis text is seen as written descriptions of historical monuments, stored in a table with many other similar descriptions in a database.

Although text is a valuable form of data, the use of photographs/paintings (images) may also be looked upon as a vital form of data. In short, an image, be it digital or not, is a visual representation of a real-life entity or entities, produced on a medium. Such a medium could be on paper as a painting, on photopaper as a photograph or digitally as byte that represent pixels on a computer. In relation to what a person may find most important, or what is in focus in a picture, it seems safe to assume that what one person may find important or notice first in an image might be considered less important by another person.

In addition to text and image, a third form of data, audio, exists. In short, audio is simply the broadcasting, reception, or reproduction of sound. Examples of audio can be the real-time broadcasting of music, a recording or simply a person talking or singing. However, in this thesis, audio refers to stored audio files on a computer or a mobile phone. These audio files contain identical descriptions of the text describing historical monuments mentioned above.

Sections 2.3 and 2.4 will go into greater detail on how these three forms of data are, or may be, used by users with regards to this thesis.

2.2 REPRESENTING INFORMATION

An important concept related to the representation of information is data. As presented in section 2.1, data can come in many different forms. An example can be a 6-digit number, 120585. With no context around the number, it may be difficult to understand what the meaning of it is. However, in the context of dates it would make sense that the number actually represents a specific date; May 12th, 1985. In other words, context around data is necessary to be able to extract information from it. Once (informative) data is gathered, it may be presented to users in various ways.

In the case of *text*, information could be represented in the form of a receipt from a purchase of some sort, as a menu in a restaurant, or even as the text in a book. For *images*, ways of representing information may possibly be obtained by "reading" the information in the actual image. This simply means looking at the contents of the image. By taking a photograph with a camera one may develop pictures on paper and place them in albums, but there are other methods commonly used nowadays. One such method is to digitally scan an image produced on paper, so that the image gets stored on a computer. Although this may still be done by some, the introduction of digital cameras has improved this process a great deal. Instead of scanning an image, one can simply upload digital images taken with a digital camera to a computer. How digital images may be used will be explained later in section 2.4.

Finally, it was given examples earlier that *audio data* can come in the form of a recording, or simply as someone talking. From both of these forms, more forms of representation can be

derived. The example of recording can be split into two categories; recordings of human speech and recordings of synthetic speech (Gong & Lai, 2001). As for talking, this form of audio data can be presented by people talking “live”, or face- to-face, as well as through a recording. These two categories are presented in more detail in section 2.3.1.

2.2.1 CROSS-MODAL INFORMATION RETRIEVAL

Based upon the various ways in how information can be represented, the process of how (electronically) searching for information can be performed will now be discussed. Possibly the most basic way of searching for information is known as text-based information retrieval (TBIR) using text as search terms for searching for information. In the early years of electronic retrieval it seems safe to assume that this was primarily executed on text collections. However, in more recent years this has evolved considerably, expanding to not only text collections, but also images, audio and video collections. With so many ways of searching among the different forms of data people are now, and have been for a long time, able to search for a certain form of data using a different form of data than they are trying to find. Put in another way, they are able to perform cross-modal information retrieval, or CMIR, which simply refers to the process of using one media form (for example text) for querying information and being return with results in another media form (for example video) (Owen et al., 1999).

As opposed to searching for text with text, a simple example of CMIR is searching for, say, images on the Internet using text as search terms. Another similar example of CMIR can be found amongst the millions of users of www.youtube.com where, again, text is used as search-terms for searching for video clips. However, for these types of searches to be performed it is necessary that all image and/or video material be annotated with text descriptors. In other words, although a search with text is performed to retrieve data in a different form, the search simply finds images or videos associated with the text annotations. Alternatively, from using text as search-terms one may also go over to searching for audio data utilizing images as “search-terms”. Thinking back to the introductory scenario in chapter 1 with tourists wanting to look up information on historical monuments, this is a clear example as to how CMIR and TTS may be connected with regards to possible, useful, uses of TTS in a real-life situation.

Now for a closer look into the various ways of conducting CMIR. The first form of CMIR to be brought to light is text-to-image retrieval. Text-to-image retrieval is presumably the most basic, and earliest, form of CMIR. In order to search for images with text it is necessary to annotate the images (add references to specific keywords) so that people searching for images connected to those words as search terms will be able to find them. This form of CMIR is also known as annotation-based information retrieval (ABIR), querying a (multimedia) database by referring to annotations connected to specific data items. ABIR may be said to be the most commonly used approach for Internet-access to image databases by utilizing "keyword matches". This simply means matching keywords given by users with annotations manually defined for each image in a database (Elgesem & Nordbotten, 2007). There is, however, a problem with this, being that not every person will describe an image in the same way. If an image is searched for, and the search term given does not match any of the corresponding annotations, it may prove difficult to retrieve the image(s). These examples also go for the CMIR form text-to-video.

Another form of CMIR, mentioned above, is image-to-audio. An example of such a search can be using an image of a real life entity and trying to find audio clips possibly related to what is in the image. This form of search may also be known as Query-by-Example or Query-by-Context which fall under the retrieval form content-based image retrieval (CBIR). In short, CBIR involves the use of different structural characteristics of an image (shape, color, texture and other factors) for searching and retrieving similar images. These factors play a vital role for searching and retrieving information by CBIR. It is also fully possible to combine CBIR and ABIR/TBIR to try and improve the quality of the results, however, in some situations this may not be necessary. An example where combining both forms of retrieval may prove pointless could be found when searching for images in a collection of images consisting of the same type of images, for example finger prints. Text identifiers would have little, or no effect to the quality of the results found. This, however, is not the case when a search is to be performed on a multimedia database consisting of a variety of images (people, animals, nature, and so on). The images most important, or in focus, for a search need to be filtered out from the rest of the "unwanted" images to further narrow down what is most important.

There are however flaws to CBIR such as the one presented in the following quote:

> *"[…] This approach suffers from the gap between the user's understanding of the semantic meaning of a search image and the current inability of the image retrieval algorithms to identify objects within the image and thus recognize its semantic meaning." (Elgesem & Nordbotten, 2007)*

Thinking back to the introductory scenario of tourists searching for information, there exist several ways in which this can be done. Of course a tourist could just find a book on historical monuments and read from it while looking at a specific monument, however, there may exist better solutions. A possible better solution introduces yet another form of CMIR; text/image-to-audio. Starting with text-to-audio retrieval, assuming tourists knew what they were looking for, one could simply search for audio clips describing the specific landmark. With image-to-audio[7] let's assume that they are willing to use their mobile phones while sightseeing. It could then be possible to take a picture with their mobile phone and use that picture as a search term to find audio clips connected to it.

2.3 TEXT-TO-SPEECH

Text-to speech, referred to as "TTS" from here on out is the artificial production of human speech utilizing a specially designed speech engines (software systems). By inserting a body of text into such an engine, words are processed and the corresponding stored audio for each word is strung together to form full sentences identical to what was first written.

When considering the quality of the output from TTS conversion, two important aspects are considered: naturalness and intelligibility. Naturalness describes how closely the output sounds like human speech, whereas intelligibility refers to the ease with which the output is understood (Acapela-Group, 2009). In other words, when talking about the quality of synthetic speech these two aspects play a vital role. With regards to naturalness, if the product of a TTS engine results in choppy and poorly strung together words this can be seen as poor naturalness in the speech as it sounds nothing, or at least little, like human speech.

[7] At the end of this thesis, in section 5.3 - FUTURE RESEARCH, a prototype is proposed concentrating on the image/text-to-audio form of CMIR.

Other issues such as pitch and pronunciation may vary in a single audio clip, and can also be seen as poor naturalness. It may be reasonable to assume that if the naturalness of synthetic speech is poor, the intelligibility of the speech will also be poor resulting in audio clips of synthetic speech being difficult to understand. Although naturalness and intelligibility can be looked upon as quality *aspects*, another way of looking at them is as quality *goals*. Any developer of a TTS engine would most likely have these aspects as main goals towards the production of better synthetic speech. It does not seem likely that anyone would purposely set these goals low, resulting in poor and hard to understand speech. With regards to what these aspects imply (more human sounding and easy to understand) it seems highly accurate to assume that any developer of synthetic speech would set these goals as high as possible to achieve the best possible outcome, or product, of synthetic speech. As to how these goals can be achieved, testing and analysis presented in chapters 4 and 5 will bring to light some of the ways in which naturalness and intelligibility may be improved, among other things

Two main concepts in relation to TTS are those of *text* and *speech*. Text could possibly be said to be the fundamental building-block of TTS and is used as input data in a TTS engine. The engine will then create an audio representation of the text. In other words, a TTS engine is simply a system used for actually converting text to audio clips. Although early versions of TTS techniques resulted in rather robotic sounding speech, over time, and with better equipment (hardware) and software algorithms, products of TTS have become increasingly more natural sounding.

As for the second fundamental concept to TTS, speech, as mentioned in an early article on TTS, despite it *"[…] being performed, remarkably well, by a number of laboratory systems and commercial devices"* (Klatt, 1987) TTS is still a ways away from being flawless. Although many TTS engines which are up-and-running exist, it is safe to assume that most of them share a common problem, pronunciation. During the process of converting text to audio, words are picked from a collection, or database, of recorded words. It is here the problem of pronunciation occurs resulting in unnatural sounding speech (Klatt, 1987). In natural speech the pronunciation of a single word will differ depending on the context it is said in. A simple example of this is the word "mine". As a statement the word "mine" may be pronounced in a determined high tone, whereas if said in an answer to a question the word "mine" can be

pronounced in a soft and lower tone. Another example of this may be seen in the following figure:

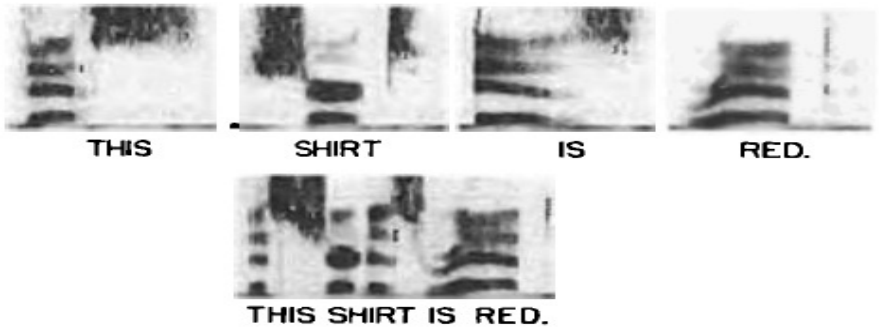

FIG. 24. Broadband spectrograms indicating that a sentence is very different from a concatenated string of words recorded in isolation. Words in sentence context are generally much shorter in duration, are subject to coarticulation at word boundaries, and undergo phonetic recoding—for example, the /t/ in "shirt" has become a flap.

Figure 2.1 – Stringed words vs. normal/actual sentence (extracted from (Klatt, 1987))

This figure is another example of how the pronunciation of words may differ based on the manner in which words in a sentence are presented. In other words, the presentation may determine the choppiness and naturalness of sentences.

A solution to this problem could be to record every variant of every word and store it into a database, but this is highly unlikely to happen. Two reasons for this are size and function. The size of a database in a case like this could end up being extremely large, and it may seem rather difficult, if not impossible, for the functionality to be able to pick out which version of a word to be used. When trying to achieve a more natural sounding speech, it has been said that stringing together stored words is simply not recommended. Factors such as timing, intonation and allophonic detail play an important role in the naturalness of the speech generated by TTS engines, and are factors hard to implement in rules and code for an engine to pick up on.

2.3.1 ACCEPTANCE STUDIES

There has been much research done on the field of TTS, dating back to at least 1987 (an example of these being (Klatt, 1987)). However, over the years many articles have revolved

around the *technical* side of how conversion rules and tools are built. Although important to TTS, they are not directly relevant for this thesis.

The research that lies closest to the work conducted here, studies questions of acceptance and behavior by users while using a system where synthesized speech plays a major role. An example of this can be found from Nass & Min Lee (2000) on whether or not synthesized speech can have "personality" and thereby influence users in different ways when using a system. Their study examined if users (72 total, 36 extravert and introvert) responded to personality cues in synthetic speech similarly to human speech. Their testing was conducted by having participants listen to audio reviews of five books in a 2x2 experiment ((synthesized voice personality: extrovert vs. introvert) x (participant personality: extrovert vs. introvert)). In Addition to listening to the audio reviews, participants were asked to fill out a questionnaire (on the writer of the reviews and the voice used for playing these reviews).
A quick summary of their findings was that participants with the same "personality" as the synthetic voice gave more positive responses. These responses were seen as participants liking the voice more, feeling more comfortable listening to someone/something somewhat similar to themselves (extrovert/introvert), and seeming more willing to purchasing the books they heard reviews of.

Although what was being investigated differed (personality vs. voice quality), Nass, Lee and Braves' works (2000) relates to this thesis in that many of the methods used for testing were similar (participants listening to clips of synthetic speech and have time to react followed by a questionnaire to reflect on these reactions). With regards to the 2x2 experiment, because this thesis presented the assumption of human over synthetic speech being preferred by users a 2x2 experiment was not needed as there was only one criterion in focus, synthetic speech. In addition to how the experiments were conducted, results as to how users responded to audios of synthetic speech are somewhat helpful in this work in the sense that it might give a more wholesome idea as to how users respond, not only with regards to personality or gender, but also simply to the quality of the speech.

A second area of research on TTS, from Nass, Lee & Brave (2000) is very similar to the latter, but differs by examining whether or not gender can influence users in how they may use a

system. In their article they mention how they used a total of 48 participants (24 male and 24 female) for testing, where they performed in 2x2 experiments and later were compared to one another for conclusion. In these cases the 2x2 experiments were in the form of ((synthesized speech: male vs. female) vs. (human speech: male vs. female)), presenting users with social dilemma situations and trying to "persuade" users into choosing one of two options. In addition to listening to the reasoning, from the computer, of which options participants should choose, users also had to answer a pen-and-paper questionnaire. It was discovered that the male-voiced computer exerted greater influence on the participants' decisions in comparison to the female-voiced computer. Furthermore, they found that gendered synthesized speech triggered social identification processes (female participants feeling more comfortable listening to female synthetic speech and likewise for male participants and male synthetic speech). Nass, Lee and Braves' (2000) work does, to some extent, relate to the work done in this thesis. In both cases aspects of synthetic speech are being testing and evaluated to see how participants and users react when presented with such speech, as well as determining what effects users most with regards to their choices made. With regards to the methods used for testing, in the case of Nass, Lee and Brave utilizing a 2x2 experiment is understandable given they wished to compare results from the computers' male and female voices with results from human male and female voices. In the case of this thesis, given it was the *quality* of the speech that was to be examined It did not seem necessary to conduct a 2x2 experiment based on the assumption that people, in most cases, would rather listen to an actual person talking rather than a computer. Although these two examples are relevant to some extent, this kind of research focused on whether or not different aspects of TTS-produced audios, such as personality and gender, can influence the way users react. In contrast, this thesis focuses solely on the *quality* of the synthetic speech and how the audios are presented to users (small, medium, long clips). By inviting participants to listen to audio clips of synthetic speech followed by answering a questionnaire of their reactions to the quality (choppiness, understandability, and so on) of the clips, aspects as to how user acceptance to synthetic speech will be investigated in this work.

2.3.2 PRESENTING SYNTHETIC SPEECH

When discussing audio in section 2.1 and 2.2, it was brought to light how audio can be presented or, with regards to TTS, how text can be represented as audio. With TTS the audio

generated from converting a written text can be looked upon as a form of recording. However, the difference here compared to the recording of an actual human is the presentation of a "new" form of speech which can be considered as rather computerized. This computerized voice, from here on referred to as synthetic speech, can be somewhat unnatural to listen to with regards to pronunciation, punctuations and so on, as opposed to a *human speech*. In other words, when listening to synthetic speech one is presented with sentences stringed together from single words recorded and stored in a database. When listening to a sentence built in this manner the flow of words may be somewhat odd. Respectively, when listening to (normal) human speech it would sound like any person talking or singing with an even (normal) flow of words, punctuation and so on.

Looking into the quality of today's synthetic speech, there has been considerable improvements in its development. Examples of such an improvement can be found in an article on the status of a trainable TTS system at IBM (Eide et al., 2003) and, in even more recent years, in an article on IBM's voice conversion systems (Zhiwei et al., 2008). Zhiewei et al. discuss recent changes to TTS algorithms as well as explain that by introducing a new "speaker", changing the speaking style and increasing the amount of recorded data used for converting text, they experience a significant improvement to the intelligibility, naturalness, prosody (patterns of stress and intonation in a language), and social impression of their female voice.

2.4 CHAPTER SUMMARY

Chapter 2 has presented the theoretical framework surrounding this thesis, and shed light on the primary topics of interest such as data forms, information retrieval and presentation, text-to-speech, and cross-modal information retrieval. The different forms of data, such as text, images and audio, were presented, as well as the various ways in which retrieval can be conducted with these forms of data (text on text, text on image, image on image, and so on). The introduction to text-to-speech presented pronunciation as a common problem amongst TTS engines, resulting in an unnatural form of speech. However, the quality of synthetic speech keeps improving, and it was also mentioned other benefits, such as the aspect of size, which make TTS systems desirable. Furthermore, a short overview of earlier research was given, and similarities between those and this thesis have been mentioned such as how

testing has been conducted earlier and how this, to some extent, matches the framework used in this thesis with regards to participants first listening to samples of synthetic speech and followed by the answering of a questionnaire.

3 DATA COLLECTING FRAMEWORK

This chapter presents the framework used for preparing the data collection phase of this thesis. In section 3.1 an explanation as to how audio clips were generated and segmented is presented. Section 3.2 looks into the various users, as well as testers, of synthetic speech, explaining how participants were gathered for conducting the test sessions, as well as presenting various attributes of the participants such as age and educational level. Finally, section 3.3 gives an overview over the similarities and differences of the two test sessions used in this thesis, and will explains how these were conducted.

3.1 CONSTRUCTING AUDIO CLIPS

The first step needed for testing was to gather text descriptions for some of the many monuments in the city Bergen, Norway. The texts used were collected from a database, BergenBy, developed by the CAIM (Context Aware Image Management)[8] research group (Langøy, 2008). The BergenBy database contains images and texts of historical monuments in Bergen. Once a set of texts, selected with no real criterion as they were all descriptive descriptions of monuments, was extracted from the database it was necessary to translate some of them as many were originally written in Norwegian. If not translated to English this would have caused a problem with conversion as the TTS engine used in this thesis is designed to be used on English texts.

When all the texts were translated to English they were converted into audio files. This task was conducted by the *Speak Computer*[9] text-to-speech engine. When deciding which TTS engine to actually use for this purpose, a minor analysis and evaluation was performed. This analysis and evaluation was based on the fact that many tools for this task exist today, but with a difference in results between them. While deciding which tool to use, focus was put on whether or not the tool was freely available, how easy it was to use, whether it was possible to save converted text as individual audio files, and the quality of the output from the tools.

[8] http://caim.uib.no
[9] SpeakComputers' homepage.

After downloading and testing several TTS engines[10], SpeakComputer was finally chosen to be used for generating audio clips for this thesis.

The process of this engine was straight forward, and is explained below:

Firstly, converting the texts involved writing (or simply copy+pasting) a specific text into the textbox of the TTS engine.

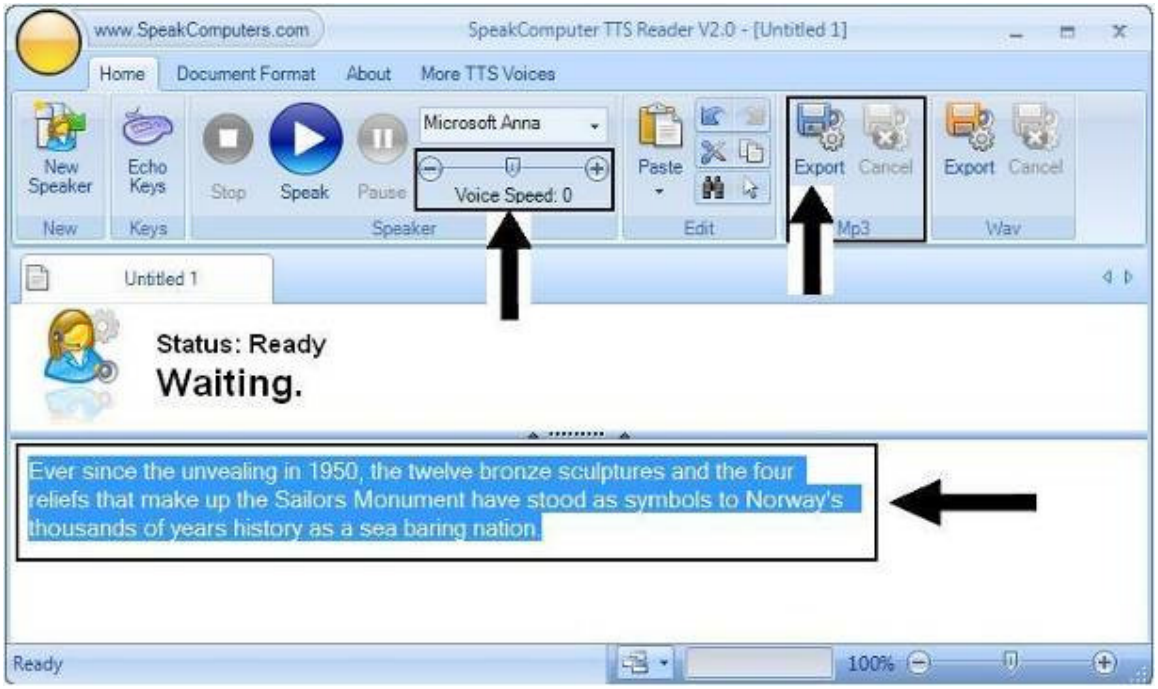

Figure 3.1 - Screen shot of TTS engine while converting a text to audio

Once the text was written (indicated by the marked text in figure 3.1) it was necessary to define the speed of the speech (indicated by the arrow pointing at "Voice Speed"). As there was no general definition given as to how fast the speech in the audios should be, it was decided to set the speed to the default setting of the TTS engine used (zero-value). To check how this value affected the speed, the higher the vale the faster the speech, and respectively the lower the value the slower. The TTS engine had two formats for how the audio file was to be saved as, .MP3 or .WAV. By clicking the "Export" button above the desired format (in this

[10] Respectively NaturalSoft's homepage, Sayvoice's homepage, ReadPleases homepage and Tanseon System's homepage as mentioned in chapter 1, section 1,2. Actual links are listed in the bibliography under "Internet Sources".

case the upper right arrow next to “Export” and “Mp3”), a “save as” dialog would be prompted. Once the location for where the audio should be saved was decided, the conversion would begin.

When deciding what format to save the audio clips as, it was first necessary to investigate the differences between saving the audio as a .MP3 or .WAV file. After saving several different texts in both formats it was revealed that the only real difference was the actual size of the files. It was found that files in .WAV format were, in several cases, 10 times the size of the exact same file in .MP3 format. There was otherwise no audible difference in the quality of the generated audios. The decision finally came down to saving the audios as MP3 files due to the fact that when conducting the mobile-based testing the audio clips needed to be saved on the actual mobile phone. The memory space on the phone was somewhat limited, so saving the clips in MP3 format ensured that it would be enough space for all the clips on the phone.

Once the audio clips were generated, the topic of length, or duration, of the clips became an issue to solve before actually saving anything onto the mobile phone used for testing. Suggestions were derived as to whether or not clips should be segmented into several shorter clips or kept as a single clip. As this thesis is trying to discover how synthetic speech can become more accepted by users, it was later decided to keep all audios as single clips. By doing so it was then possible to test clips of short, medium and long durations on users to see if duration actually plays a role in acceptance. The following explains how the different groups of audio clips were derived.

When testing was to be conducted it was decided to divide a selected set of audio clips into three groups. These groups were defined as clips of a specific duration; short, medium and long. Upon dividing the audio clips it was first necessary to determine each group’s duration. The selected set of audios to be used for testing varied from clips of 7 seconds to clips of 134 seconds. Defining the duration of each group of clips was based on several factors. Firstly, Müller et. al. (2005) defined the length of a “short audio fragment” to be between 10 and 30 seconds. As to how they came across these numbers there is no explanation given. Secondly, the number of generated audio clips, as well as their actual duration, played an important role in defining the small, medium and long groups’ durations. In total there were 20 unique audio

clips stored with durations ranging from 7 to 134 seconds[11]. A list of these audio clips and their durations is given below:

AUDIO ID	AUDIO TITLE	DURATION IN SECONDS
14	Rogalandskrosset	7
16	Sjømannsmonumentet	11
6	Festplassen	15
3	Børs	21
5	Edvard Grieg Statue	24
17	Statsraad lehmkuhl	24
8	Gamlehaugen	26
13	Ludvig Holberg Statue	35
20	Troldhaugen	39
10	Johanneskirke	40
9	Grieghallen	41
11	lille lunggårdsvannet	41
15	Rosenkrantztårnet	44
4	Christian Michelsen Statue	46
1	Beffen	49
19	Sæverudmonumentet	83
2	Bergen Natural History Museum	89
7	Fisketorget	94
12	Long church	120
18	Stavkirke	134

Figure 3.2 - Table of existing audio clips (by ID), title and their durations (in seconds)

In sorting the durations from lowest to highest it was possible to divide the audios into 3 groups defined as short, medium and long. Given the number of clips (20) it was not possible to split them all into groups of even numbers. Alternatively, it was decided to split the clips into as even groups as possible. This was to attempt to have an equal number of clips to

[11] A more detailed list of what each clip described is presented in appendix D with both title and corresponding ID from figure 3.2, as well as the text converted for making the audio.

choose from while creating the pre-selected sets presented to participants during testing. The result was a 6-7-7 split. The results are depicted in figure 3.2 using a dotted line to divide the three groups:

Short: 0-25 seconds
Medium: 26-45 seconds
Long: 46 seconds or greater

3.2 USERS OF SYNTHETIC SPEECH

3.2.1 USERS

Users of synthetic speech can be found in a variety of situations; tourists walking in a city, people listening to a GPS devise or even people with certain disabilities. The range of potential users for a system utilizing synthetic speech may vary in a number of manners. The introductory case in chapter 1 focused on the tourist aspect of searching for information (on landmarks), where the user group for synthetic speech consisted of both local and foreign tourists. An assumption made of tourists is that systems utilizing synthetic speech would be used to a larger extent by locals rather than foreigners due to prices of Internet services while abroad. This, however, is a changing issue due to such prices, presumably around the world, becoming cheaper and therefore making it easier for foreign tourists to use their own mobile devices for searching for information.

Tourists are only one of the possible groups of users that can draw benefits from the use of synthetic speech. Another group of users are those with disabilities of some sort. These users may still be tourists, but this specific “kind” of tourist gives a vantage point to the more practical and social/personal benefits of using such speech. Instead of having to read text on a mobile phone’s display, it may prove more satisfying for a user to only have to find the information they are seeking and simply have the system/mobile telephone “read” it out for them.

Through discussion with several different, possible, users about possibly using a system that utilizes synthetic speech, many commented on certain issues both positive and negative. Among these it was found that more elderly people seemed to like the idea of not having to

read text on a small display, and instead having someone else reading the text to them. Furthermore, younger people also agreed that having a computer voice reading to them seemed a very positive aspect of such a system.

Other issues brought to attention in these discussions are on the different varieties of how information may be given, or presented, to a user. Comments such as “I prefer both text and sound at the same time” and “why not use human speech” were some of the comments mentioned the most frequently. Upon playing a recording of synthesized speech, it seemed that the people listening didn’t mind the choppiness, or lack of naturalness, in the voice as much as they first thought they would and thereby, rather quickly, accepted this form of speech. However, this was only a brief observation from a few people, and should not be considered a form of conclusion. In chapter 5, results from the analysis of the collected data from testing will be presented, followed by possible conclusions based on these findings.

3.2.2 TESTERS

The people who were invited to test the clips of synthetic speech were randomly picked by asking first year students from the University of Bergen (UiB) if they were interested. The reason for this choice in participants was based on the fact that anyone could be a potential user of a system utilizing synthetic speech for informational purposes. “Anyone” primarily refers to the age of participants. A tourist, for example, can be a young teenager being 18 years old, or it could be a grown person well into his/her 40’s, 50’s, 60’s or older, in other words “anyone”. This is also true for a person with disabilities, or for someone using a GPS device. It is not that uncommon to find students, in the same class and year, with large age differences between them. As it will be presented in chapter 5, when finding participants for testing in this thesis, they varied in age from 20 years old to over 40 years old. Because of this, using students to represent possible users seemed appropriate and also made the test sessions easier to conduct.

As mentioned above, all participants of the testing were found around the campus grounds of the UiB. Although it was hoped to only use first year students as participants for testing, it was later decided to invite students of any year due to a lack of accepted invitations from the first year students. The original plan of solely using first year students was an attempt to limit the

possible knowledge participants have of synthetic speech. With the revised plan of using any student it was assumed that, although there might be a higher chance of participants having knowledge of synthetic speech, participants having *some* knowledge might not harm the overall results. This was assumed due to the fact that many future users may also have knowledge of synthetic speech, for example tourists in the sense that they might be familiar with GPS systems.

From the beginning stages of planning how testing would be conducted, it was always a goal to have an even number of male and female participants to keep the results gathered as representative as possible. After all testing was completed the final number of participants was 25, with 12 of them being male (48 %) and 13 being female (52 %). As to how many participants were used in each of the test session, chapter 5 will go into greater detail on this.

3.2.3 AGE

The reason for registering the participants' age was simply to investigate if this factor played any role in how participants responded to the different quality aspects of synthetic speech. For example, it was desired to find out whether a 30 year old participant would answer significantly different than an 18 year old participant to the same question. This also goes for participants of somewhat equal or identical age and seeing if any major difference occur here as well.

Because it was students that were to be used for testing, the youngest participants would be no younger than 18 years old, but it was more difficult to make any assumptions about the upper age limit. When testing finally began, the upper age limit was originally set lower, but when confronted with a participant that did not fit in any of the given options it was necessary to change this upper limit. Wanting to keep an even break in the age groups, the upper limit was finally set to participants of ages 43 to 47. Once testing finished the final limits of ages were between 18 and 47 years of age; 18 years of age falling on those who were in their first semester of university and up to 47 years for those presumably going back to university. To avoid very comprehensive figures, bulks of 5 years were used. When testing was concluded it was revealed that 56% (14 out of 25) of the participants fell within the age range

of 23-27, with a few exceptions naturally, not surprising due to participants being students. Results from such investigation will be presented in chapter 4.

3.2.4 EDUCATIONAL LEVEL OF PARTICIPANTS

Following the same reasoning as for registering the age of the participants, registering the educational level was done to find out if this factor played any role in how participants responded to the different quality aspects of synthetic speech. In other words, it was to see if participants in, for example, their 4th year at University responded significantly different to participants in, for example, their 1st year.

Due to participants being found around the campus grounds, the educational level of participants seemed to vary from students either in their early or later years of bachelor degrees to likewise for master students. The exact numbers gathered from the 25 participants are represented in figure 3.3:

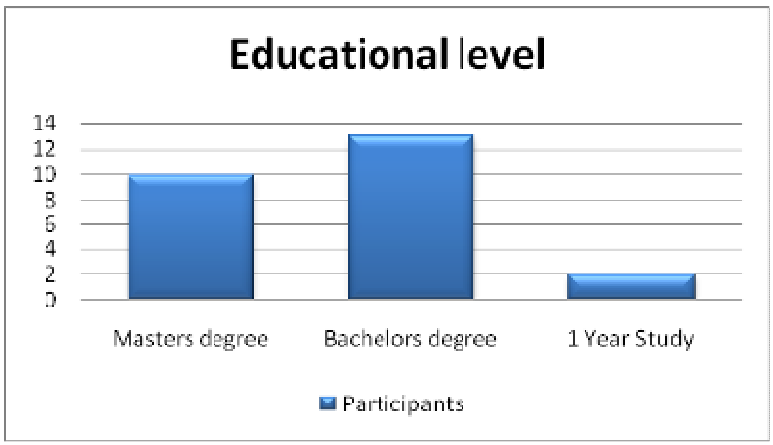

Figure 3.3 – Diagram of educational level of participants from testing

3.3 EXPERIMENT FRAMEWORK

When planning how the testing phase of the thesis would be conducted, only one round of testing was initially envisioned. The testing was to be done by combining a laboratory study and questionnaire. This plan was later revised and an additional testing round was added; a field study, also combined with the same questionnaire. The reason for adding the second test session was to investigate whether or not being in a different environment would affect

the participants' responses in the questionnaire. Put differently, would the responses of the participants in the laboratory study and the field study turn out to be significantly different? If not, this would indicate that all the participants could be combined into only one testing group.

The two test sessions were later called mobile-testing (real-scenario test) and computer-testing (the in-house test) due to the nature of how the tests were conducted.

3.3.1 SIMILARITIES IN TEST SESSIONS

Preparations for the two test sessions involved many of the same activities and tasks. To begin with, it was necessary to inform all participants of the nature of the testing they were a part of. To be sure all participants had the same knowledge of what was going to happen, an information sheet was sent/given to the participants either before they came or when they arrived to start testing. The information sheet explained how things were going to be done, what was needed from the participants, equipment to be used, how long the testing process would take and other information on the handling of results/gathered information. A copy of the information sheet can be found in appendix A. In addition to the information sheet, participants were informed that if they had any questions during the process of testing they should ask whenever they wanted to.

Once participants had read the information and signed a "Participant Consent Form", shown in appendix B, stating that they approved and agreed to the proposed rules, the next stage was to explain what they were to do from there. At this point the list of tasks to perform during the testing varied slightly between the two test sessions. These differences will be explained in section 3.3.2. What was still the same in both sessions was the fact that participants were to first listen to audio clips from a specific duration category, and then complete a questionnaire. Depending on the order of duration category ("long →medium →short" or "short →medium → long") participants would, for example, first listen to two clips from the "long" category followed by the questionnaire. Once they had done this they would then listen to two more clips, but this time of clips from the "medium" category, and again followed by the same questionnaire as before. When participants had completed this

they would move to the final set of audio clips, from the “short” category followed, for the last time, by the questionnaire.

The tools/equipment used during testing, in both sessions required participants to use headphones to listen to the audio files. The reason for this was based on the fact that it seemed highly unlikely that users would listen to audio descriptions while having to hold a mobile phone to their ear while walking around in a large city sightseeing. It was concluded through general conversation that the use of headphones would make hearing the audio descriptions easier based on the fact that headphones could block out more of the surrounding noise caused from other people, cars, animals, and so on around the participants. In order to get a feeling for the real-life situation, the participants on the computers were also equipped with headphones. Other than the use of headphones the only other equipment was the actual tool used for playing the audio files. In the first test session testers were asked to use a laptop computer, whereas in the second session a mobile telephone was used.

3.3.2 DIFFERENCES IN TEST SESSIONS

In figure 3.4 a list is presented of all the audio files that were planned to be used for testing, and the sets in which they were placed into.

Duration Category	Set 1	Set 2	Set 3	Set 4	Set 5
Long	Fisketorget	Christ. Michel. Statue	Beffen	Fisketorget	Bergen Nat. Hist. Museum
Long	Christ. Michel. Statue	Bergen Nat. Hist. Museum	Sæverud Monument	Beffen	Sæverud Monument
Medium	Rosenkrantz Tower	Grieghallen	Troldhaugen	Ludvig Holberg Statue	Gamlehaugen
Medium	Gamlehaugen	Ludvig Holberg Statue	Grieghallen	Rosenkrantz Tower	Troldhaugen
Short	Edvard Grieg Statue	Sailors' Monument	Børs	Rogalnads' Cross	Edvard Grieg Statue
Short	Rogalands' Cross	Statsråd Lehmkul	Sailors' Monument	Statsråd Lehmkul	Børs

Figure 3.4 – List of pre-selected audio sets used during testing

When putting together these sets, two criteria were considered; one: that each set had two clips from each of the three duration categories, and two: that the sets would vary as much as possible with regards to the clips used in each set. To the second criteria, this simply involved

making sure certain clips were not used in more sets than others, keeping an even balance amongst the clips. As can be found in figure 3.5 no clip was used in more, or less, than 2 sets. Otherwise, as all the clips were of various monuments there was no criterion as to what the clips contained, be it a description of a boat, a building or a statue. To the left in figure 3.5, under "duration category", are the categories in which audio files were grouped into. On the right side in the figure are the names of the place, person or monument the different audio files describe. Due to fewer participants than planned, some of the audio sets needed to be dropped while conducting both test sessions. The differences lie in which clip sets were dropped. In the computer-testing, clip set 5 was dropped, while the mobile-testing only used sets 1 and 2.

There are also differences in how the test sessions were conducted. First, the mobile-based testing is presented. This session involved the use of a mobile telephone as a tool/medium for listening to audio files for specific monuments. The audio clips were saved on the mobile phone as .MP3 files and were listened to in the same way a mobile phone plays music files. The whole process of the mobile-based testing is explained below:

1. participants were first given an information sheet, participant consent form and set of questionnaires to read through, sign and, if need be, ask questions about before the actual testing began
2. a list of monuments, and order in which they are to be looked at (categorized by duration), was given to each participant and then they began walking around downtown Bergen
3. upon seeing a monument on the list participants took out a mobile phone and:
    a. looked through stored audio clips on the phone until they found the correct clip (defined by its filename)
    b. attached a pair of headphones to the mobile phone
    c. began listening to the audio description of the monument they were looking at
4. when participants had found and listened to two clips of a duration (long, medium, short) they were asked to answer a questionnaire[12]

[12] The questionnaire is presented in appendix C.

5. once the questionnaire was answered participants found and listened to clips of the next duration category
6. Finally, once the questionnaire had been filled out three times a final round of questioning and discussion was performed to gather any additional information and comments.

Task 4 basically repeated tasks 3 until all clips from all duration categories had been listened to.

The computer-based testing involved using a laptop computer as a tool/medium for listening to audio files for specific monuments. In this session the audio clips were stored on the computer and presented to the participants in a "playlist" for a standard media player. Each clips' filename consisted of a category (long, medium, short) and the name of the monument, for example "Long-Fisketorget" so that participants were aware of the differences of the clips beforehand. The testing was conducted in a small room with only the participant and the researcher, a computer and a pair of headphones. The testing process is explained below:

1. Participants were given an information sheet, a participant consent form and set of questionnaires to read through, sign and, if need be, ask question before the actual testing began
2. A pre-selected set of audio clips was presented to the participants in a playlist sorted by duration category
3. Participants put headphones on and started listen to the audio clips in the presented order
4. When two clips of a certain category had been listened to participants were asked to answer a questionnaire
5. When the questionnaire was filled out participants would then continue listening to clips of the next duration category until all clips were listened to and the questionnaire was filled out three times (once for each duration category)
6. Finally, once the questionnaire had been filled out three times a final round of questioning and discussion was performed to gather any additional information and comments.

Task 5 here simply repeated tasks 3 and 4 until testing was complete.

As the two process lists for the different test sessions show, the main differences are found in steps 2 and 3. In other words, the first difference is *where* the testing was conducted. During the mobile-based testing participants were asked to walk around Bergen, whereas during the computer-based testing they were sitting in a room in front of a computer the whole time. The second difference is how the participants were *presented* with the audio clips they had to listen to. During the mobile-based testing participants were presented with a list clips to listen to and in what order they were to be listened to. For every clip listed, participants had to walk to that monument and find the corresponding clip stored on the mobile phone. Meanwhile, for the participants in the computer-based testing, the pre-selected clips were already in a playlist in the correct order.

Chapter 5 will present all the results of the collected data from both test sessions, computer-based and mobile-based testing, as well as analyze these results.

3.4 CHAPTER SUMMARY

Chapter 3 has presented the main issues of importance around the data collection stage of this thesis. First and foremost, the generation of audio clips was presented. Here it was explained that texts first needed to be extracted from an existing database, followed by translating many of the extracted texts as they were primarily stored in Norwegian. Once this was done a TTS engine, SpeakComputer, was used to finally convert the texts to audio clips in MP3-format. Following this was the issue of segmentation, which involved organizing the clips into different groups of durations (short, medium and long clips). Following this, issues surrounding users and testers have been presented to give an idea of what kind of people may be potential users of synthetic speech. Tourists and people with reading disabilities were mentioned as examples of this. Furthermore, the lack of first year students was explained and the inviting of students from all levels, adding up to a total of 25 participants. It was also presented information about the age and educational level of the participants. Finally, a presentation of the testing framework explained how the participants were given an information sheet, a consent form and a questionnaire. The testing involved the use of two

separate test sessions, mobile-based and computer-based testing, and the similarities and differences of each test session were explained.

4 ANALYSIS

Now that a framework for how the data was to be gathered has been explained in chapter 3, attention will be turned towards the actual analysis of these data. Section 4.1 and 4.2 will present the results from respectively the computer-based testing and the mobile-based testing. Section 4.3 looks at the combined summary data for *both* test sessions in order to find out whether or not the participants' responses are affected by which of the two sessions they are a part of. Finally, section 4.4 summarizes and gives an overview of the most important indications found in these two test sessions.

Before presenting the results, it might be appropriate to clarify a few of the aspects of the testing. Firstly, the four main questions that the participants needed to answer were as follows:

1. Was the speech smooth or choppy?
2. Was the speech difficult to understand (was it clear/unclear what was being said)?
3. Was this form of speech tiring to listen to?
4. Could you listen to this form of speech for a longer period of time without wanting to stop?

It is worth noticing the consistency between the *questions* chosen for the questionnaire and the actual *theory* presented in chapter 2. Thinking back to the theory about text-to-speech, the two most important aspects of a TTS system were defined as naturalness and intelligibility. Because of their importance, these two aspects are integrated into respectively question 1 and 2 listed above. The discussion of naturalness focused on how closely the synthetic voice sounded like a human voice, and this is exactly what question 1 tries to shed more light on. If the participants perceive the synthetic voice as very choppy, for example because of poor pronunciation, this could indicate a low level of naturalness. In the same manner, perceiving the speech as smooth could indicate a higher level of naturalness. Regarding the other important aspect of TTS, intelligibility, this referred to the ease with which the synthetic voice is understood. As can be seen, this aspect of TTS is directly reflected in question 2, seeing as this question is designed to measure the level of understanding amongst the participants when listening to the audio clips. If responses from the participants

show a low level of understanding, this will also indicate a low level of intelligibility and vice versa.

After listening to the audio clips, the participants needed to use a scale to rate every question. The lower the level of rating, the more negative the answer was. For example, on the first question in the questionnaire[13], there were 7 ratings to choose from, ranging from "very choppy" to "very smooth". If the participant chose "very choppy" this would be the equivalent of a 1, whereas if "very smooth" was chosen, this would be the equivalent of a 7.

When referring to the order in which audio clips were played to participants, with regards to their durations, the following abbreviations are used; L = Long, M = Medium, S = Short. Also, when referring to the term "foreign word" in the following, this doesn't mean that the word is foreign to the *participants* in the testing. On the contrary, it means that the word is foreign to the *TTS-engine*. More precisely, the word is in a different language than what the engine is originally designed to convert. In the case of this thesis, Norwegian words (for example the names of monuments), are "foreign" because the original language for the engine is English. Finally, for a detailed overview of the results from each participant refer to appendix E.

Regarding the two follow-up questions in the questionnaire (questions 5 and 6), these were mainly asked to find out the participants' suggestions as to how the synthetic speech could be *improved* and thereby also how the participants' *acceptance* of the speech would increase. Because these suggestions are directly connected to the research question of this thesis, "How can the acceptance of synthetic speech be improved?", they will be presented and evaluated in chapter 5.

4.1 COMPUTER-BASED TESTING

When the computer-based testing was completed, and results were gathered, a total of 17 participants had been recorded. Out of these there were 9 female and 8 male participants. As explained in the previous chapter, results will initially be split into 2 main groups; male and

[13] The complete questionnaire can be found in appendix C

female. Due to this the results here will be presented by first looking at the female results in section 4.1.1, followed by the male results in section 4.1.2.

4.1.1 RESULTS FROM FEMALE PARTICIPANTS

Out of the 9 female participants in the computer-based testing, 5 listened to audio clips in the order long-medium-short and the remaining 4 listened to short-medium-long. In the tables below participants in white (1, 2, 5, 8 and 15) represent those that listened to audio clips in the order L-M-S, and those marked in blue (4, 6, 13 and 16) are those that listened in the order S-M-L. This setup goes for the tables presented later in this chapter as well.

Question 1) "Was the speech smooth or choppy?"

Participants:		1			2			4			5			6			8			13			15			16		
Gender:			F			F		F				F		F				F			F			F			F	
Duration:		L	M	S	L	M	S	L	M	S	L	M	S	L	M	S	L	M	S	L	M	S	L	M	S	L	M	S
Audio Quality:																												
choppiness	levels 1-7	1	1	1	1	3	3	2	2	6	2	5	6	5	5	3	4	3	5	5	4	2	1	3	4	4	2	3

Figure 4.1 - Extract of raw data from female participants of question 1 of the questionnaire

Order: L→M→S:

Of the 5 participants, the majority (4 out of 5) responded with more positive answers the more they listened to the clips. "More positive" simply means that the level of choppiness was considered to gradually decrease during the sequence. In other words, most (4 out of 5) of the participants found the clips to be smoother at the end of the sequence than at the start. However, one participant found the speech to be equally choppy in all the clips that were played.

Order: S→M→L:

Of the 4 female participants that listened to clips in this order, the majority (3 out of 4) responded with similar results to those that listened to the L→M→S ordered clips. Results

from both orders indicate a decrease in choppiness of synthetic speech the more it is listened to. Discussions with the participants revealed that the more they were *exposed* to this kind of speech, the easier it got to understand what was being said. One participant, however, did find that the longer the clips were, the choppier the speech seemed to become, and finding medium and long clips to be equally choppy.

L→M→S and S→M→L:

Combined, the majority (7 out of 9 participants) found that the more they listened to the audio clips the less choppy they seemed to be. This can be seen as an early indication that the *duration* of clips doesn't play much of a role towards better understanding what is being said with synthetic speech. Upon discussing the gradual decrease of choppiness, participants were surprised to hear that there was actually no difference between the audio clips with regards to how they were converted from text to audio. Although the remaining 2 out of 9 participants were rather negative in their responses, it doesn't seem to impose any significant change to the indication mentioned above.

Question 2) "Was the speech difficult to understand (was it unclear what was being said)?"

Participants:		1			2			4			5			6			8			13			15			16		
Gender:			F			F			F			F			F			F			F			F			F	
Duration:		L	M	S	L	M	S	L	M	S	L	M	S	L	M	S	L	M	S	L	M	S	L	M	S	L	M	S
Audio Quality:																												
understandable	levels 1-5	3	4	3	2	2	3	4	2	4	4	4	5	4	4	4	4	4	4	4	3	4	2	4	4	4	3	2

Figure 4.2 - Extract of raw data from female participants of question 2 of the questionnaire

Order: L→M→S:

As with question 1, the majority of participants (4 out of 5) revealed either an increase in understandability and clearness of the synthetic speech during the sequence, or appeared to be consistent with their answers finding all groups of clips rather easy to understand. There was however one participant revealing a slight decrease in clearness when going from a clip of

medium duration to a short clip. This was later discussed, and the participant explained that the decrease in clearness was due to many foreign words in the short clip. Because these foreign words were being poorly pronounced, it was difficult to follow along in the text afterwards.

Order: S→M→L:

All the participants that listened to clips in this order (4 out of 4), found that the clips were relatively easy to understand. However, participants found some of the clips *slightly* more difficult to understand than other clips. Half of the participants described an increase in difficulty with clips of medium duration, but found clips of short and long duration to be easier to understand. Out of the remaining two participants, one found all clip durations to be equally easy to understand. The other one found it easier to understand the clips the more she listened to them, starting with the shortest duration to be most difficult to understand and ending with long clips being the easiest. As can be seen, the results are somewhat mixed as to what clips stood out as slightly more difficult, and may be assumed to be due to the number of foreign words in the different clips. Despite this, the main result is that all of the participants found the clips to be *relatively* easy to understand. This trend indicates that the exposure to synthetic speech improves the overall understanding of it.

L→M→S and S→M→L:

Overall, as many as 89 % of the participants (8 out of 9) found that the audio clips became either *increasingly* easy to understand throughout the sequence, or that they were *consistently* easy to understand from the start to the end of the sequence. The participants that found the synthetic speech easy to understand already from the beginning of the first clip, became accustomed to the new form of speech very quickly, while the participants that found the speech increasingly easy to listen to throughout the sequence needed some more time to adjust. In both cases, the main indication is that by being *exposed* to the synthetic speech all the participants got accustomed to it at some point, whether it was early on in the sequence or a little later. This led to a better understanding of the synthetic voice and what was being said in the clips.

Question 3: Was the speech tiring to listen to?

Participants:		1			2			4			5			6			8			13			15			16		
Gender:			F			F			F			F			F			F			F			F			F	
Duration:		L	M	S	L	M	S	L	M	S	L	M	S	L	M	S	L	M	S	L	M	S	L	M	S	L	M	S
Audio Quality:																												
tiring	levels 1-5	2	5	4	2	4	3	4	4	4	2	2	4	4	4	4	4	2	4	2	2	4	3	4	3	4	3	4

Figure 4.3 - Extract of raw data from female participants of question 3 of the questionnaire

Order: L→M→S:

Responses to this question varied considerably, making it difficult to find any indication of trends to how participants answered. While listening to the first set of clips (long), the majority of the participants (3 out of 5) chose to classify the synthetic speech as “tiring“. Out of the two remaining participants, however, one found the first set of clips to be slightly less tiring and rated them as “average”, whereas the second one was a great deal less affected by the synthetic speech and chose “not very tiring”.

Continuing on to the clips of medium duration, the results from the majority of the participants mentioned above showed signs of improvement. From long to medium clips, the rated level of tiredness decreased from ”tiring” to “not at all tiring” and “not very tiring”, which is a clear improvement. However, moving along from medium to short clips, the participants found an increase in the tiredness again, although this was only a marginal increase. As for the remaining 2 out of 5 participants, one seemed to find clips of long and medium duration to be the most tiring, but jumped from “tiring” to “not very tiring” when presented with clips of short duration. In contrast, the other one, found long clips to be “not very tiring”, but jumped up to “tiring” on medium clips, and fell back to “not very tiring” when listening to short clips.

Although it is hard to see any clear indications from these results, it is worth noticing that the majority found the long clips tiring, but that the level of tiredness decreased considerably when moving on to medium clips. Upon discussing the question of tiredness after participants

were finished answering, it was revealed that the overall decrease in tiredness was due to participants being more *accustomed* to the new style of speech combined with the shortening in duration of the clips. Some participants explained further that with improvements to synthetic speech they would most likely find the *content* of the audio clips more tiring than the voice itself, depending of course on the interests of the listener.

Order: S→M→L:

For participants listening to the audio clips in this order, the majority (3 out of 4) concluded that the clips were "not very tiring". They were rather consistent with their opinions, and the rating did not change when the duration of the clips changed. There was however one participant out of the four that started with the same positive response as the others, choosing "not very tiring", but when presented with medium and long clips the participant's ranging changed from "not very tiring" to "tiring". Overall, the average rating of tiredness for this group of participants indicate that the clips were not very tiring to listen to, with the exception of the one who found the clips to get more tiring the longer the duration.

L→M→S and S→M→L:

Comparing the results of the 9 female participants, the average response regarding the level of tiredness was rather positive with 7 out of 9 participants choosing "not very tiring" for most of the clips (with some exceptions for the long clips).

Although results are somewhat mixed for question 3, they still seem to indicate that *exposure* to synthetic speech might *decrease* the level of tiredness over time. This is confirmed by the fact that the majority of those who listened to the clips in the order long-medium-short, showed signs of a decrease in tiredness throughout the sequence. The longer they listened to the synthetic voice the more accustomed to it they became, hence getting less and less tired. The participants testing the order short-medium-long were, however, more consistent and found the synthetic speech to be "not very tiring" from the beginning.

Question 4: Could you listen to this form of speech for a longer period of time without wanting to stop?

Participants:		1			2			4			5			6			8			13			15			16		
Gender:			F			F			F			F			F			F			F			F			F	
Duration:		L	M	S	L	M	S	L	M	S	L	M	S	L	M	S	L	M	S	L	M	S	L	M	S	L	M	S
Audio Quality:																												
willingness to listen	leves 1-4	2	3	3	2	2	2	2	2	2	2	2	2	4	4	3	3	3	3	2	3	4	2	3	4	2	3	4

Figure 4.4 - Extract of raw data from female participants of question 4 of the questionnaire

Order: L→M→S:

For this question, an average between "not likely" and "maybe" was discovered.

After listening to the long clips, the majority of the participants (4 out of 5) found it "not likely" that they could listen to this form of speech for a longer period of time. Looking at the change in opinions throughout the sequence of the testing, two of these participants were rather consistent with how they felt and also answered “not likely” for the medium and short clips. The other two participants were leaning more on “maybe” after listening to the medium clips, and respectively “maybe” and “yes” for the short clips. In contrast to the majority, one participant was slightly more positive and responded with "maybe" already from the beginning with the long clips, and consistently gave this answer all the way to the end with the short clips.

The fact that the majority of participants answered maybe, or even yes, for the medium and short clips might indicate that the clips became more appealing to listen to the shorter they were. However, it might also indicate that the participants became more and more used to listening to the synthetic voice during the sequence, and that their responses therefore got increasingly positive towards the end.

From all five participants that listen to this order of clips, the reasons for their responses were based on several factors. Although the question at hand was aiming primarily at the *duration*

of the clips and not so much at the *use* of synthetic speech, most participants found that the choppiness they experienced with the speech added to the unwillingness to wanting to listen for longer periods of time. Another reason, which was not looked upon as of particular interest for this thesis, was solely based on the fact that some of the participants were not very interested in the content of the clips and therefore not interested in listening to them. This aspect will be discussed further in the conclusion of this thesis, chapter 5. Upon asking what would make participants more willing to listen to longer clips, 4 out of 5 replied that improving the quality of the synthetic speech could help, depending on the content in some cases.

Order: S→M→L:

There was no clear average response to this question. Responses varied between "yes" and "not likely" throughout the whole sequence. The majority of the participants (3 out of 4) seemed to be less and less willing to listen the longer the clips became. It was later brought to light that the use of the synthetic speech, instead of normal human speech, was understandable, but because they needed to strain themselves to listen they found that doing so for a long time was not something they were willing to do. There were, however, signs of slight improvements in the willingness to listen during the sequence. This might be because of exposure to synthetic speech, in the sense that the participants got more accustomed to the new type of speech and thereby had less trouble listening to clips of longer duration.

L→M→S and S→M→L:

Looking at the development of changes in opinion among the female participants it appeared that the majority (6 out of 9) were either willing, or maybe willing, to listen to clips longer than those of the shortest duration. Correspondingly, most of the participants also responded with "not likely" when presented with the long clips. The longer the clips became, the less willing they were to listen to even more. In contrast to the majority, there were some participants that answered "not likely" from beginning to end. They later explained that this was mainly due to either poor quality or simply just the content of the clips.

To sum up, it seems like the *duration* of the clips actually mattered to the participants when answering question 4. This is in contrast to the results for question 1, 2 and 3, where *exposure*

to synthetic speech played a large role towards affecting the participants, and duration did not matter much. However, the fact that the participants seemed to be less willing to listen to longer clips than shorter clips, only emphasizes the importance of improving the *quality* of the speech. A shared opinion amongst the majority of the participants was that improving the quality of the speech would most certainly increase their willingness to listen to this form of speech for a longer period of time. Most of the participants agreed that, looking away from the quality of the speech, 35 to 45 seconds seemed an acceptable duration for this type of speech. Clips of any longer duration than this would simply be too tiring to listen to. The female participant's suggestions as to how the quality of the speech could actually be improved are presented in chapter 6.

Now for a closer look at the responses from the male participants.

4.1.2 RESULTS FROM MALE PARTICIPANTS

Out of the 8 male participants that were used during the computer-based testing, 4 were used to listen to audio clips in the order long-medium-short and the remaining 4 listened to short-medium-long.

Question 1) Was the speech smooth or choppy?

Participants:			3		7			9			10			11			12			14			17		
Gender:			M		M			M			M			M			M			M			M		
Duration:		L	M	S	L	M	S	L	M	S	L	M	S	L	M	S	L	M	S	L	M	S	L	M	S
Audio Quality:																									
choppiness	levels 1-7	1	1	2	5	2	3	3	3	5	3	3	2	3	3	3	3	1	6	1	3	3	1	1	2

Figure 4.5 - Extract of raw data from male participants of question 1 of the questionnaire

Order: L→M→S:

There seemed to be an indication of a pattern amongst the participants that the more they listened to the clips, the less choppy they seemed to consider them. This was enforced by the majority of the participants (3 out of 4), indicating a constant decrease of choppiness, starting

with long clips being rated as the choppiest and short clips being rated as the smoothest. There was however one exception. One of the participants was consistent with his answers, finding clips of all durations to be equally choppy.

Order: S→M→L:

Out of the 4 participants that listened to clips in this order, there does not appear to be a pattern in results indicating anything as to which group of clips that seemed least choppy. Two of the participants found the short clips to be rather smooth, but then experienced an increase in choppiness to "very choppy" and "choppy" in the medium clips, only to fall back down 2-3 levels on the rating scale with the long clips and concluding that these were respectively "slightly choppy" and "slightly smooth". A possible reason for this increase could simply be due to the use of many foreign words in the medium length clips, however, as there were no comments given from these participants indicating this, the assumption is merely speculation. Out of the other two participants, one of them found that the choppiness increased during the sequence, while the other one experienced a decrease in the levels of choppiness. As can be seen, the results here are very mixed. The only sign of a possible pattern, found through conversations after testing was conducted, was that there seemed to be a slightly easier understanding of the synthetic speech after listening to many clips, however the quality was still considered the same.

L→M→S and S→M→L:

Looking at results from all 8 male participants, 75 % of them (6 out of 8) indicate a pattern of finding short clips smoother than medium or long clips. However, some of the participants found the longest clips to be the smoothest. This may simply indicate that more *exposure* to the new form of speech may lead to participants feeling like the speech gets smoother throughout the sequence, regardless of in which order the clips are listened to. Although users might "feel" a change in choppiness during the process of testing, the reality of the matter is that the clips are all produced identically. The perceived change can therefore be explained by the fact that the participants have become more accustomed to the speech.

Question 2) Was the speech difficult to understand (was it clear/unclear what was being said)?

Participants:			3		7			9			10			11			12			14			17		
Gender:			M		M			M			M			M			M			M			M		
Duration:		L	M	S	L	M	S	L	M	S	L	M	S	L	M	S	L	M	S	L	M	S	L	M	S
Audio Quality:																									
understandable	levels 1-5	2	2	2	4	3	4	3	3	4	3	3	2	2	2	2	4	3	4	2	3	3	2	2	2

Figure 4.6 - Extract of raw data from male participants of question 2 of the questionnaire

Order: L→M→S:

Half of the participants (2 out of 4) experienced an increase in understanding throughout the sequence. In other words, the more they listened to the clips, the easier it became to understand what was being said in them. Discussion around this matter showed that these participants found the synthetic speech to improve somehow during the course of the testing. They suggested, like many others have suggested in earlier sections, that *exposure* to synthetic speech could have led to a decrease in the difficulty of understanding what was being said.

The other half of the participants ranged all the three durations of the clips to be equally difficult to understand.

Order: S→M→L:

The results of the 4 participants which listened to clips in this order seem to support the indication from the previous section: by simply exposing the participants to synthetic speech, their understanding of what is being said will increase. The majority of participants (3 out of 4) found long clips to be equally clear and easy to understand as short clips, if not easier. Given that they experienced an increase in understandability while listening from short to long clips, this may be seen as an indication that duration played no role in the responses given. It was not the *duration* of the clip that affected the level of understanding. If that was the case, the shorter clips would be the easiest to understand while the longer clips would be harder to

understand. In contrast, it seemed to be *exposure* to the speech that affected how well the participants understood what was being said.

L→M→S and S→M→L:

Combing results from all the male participants, the majority of them (5 out of 8) found that over time it became easier to understand what was being said in the audio clip. Participants explained that they became more and more accustomed to the unusual pronunciation of the speaker, and therefore also more prepared for how the speaker would pronounce words later on. As for the other 3 out of 8 participants, it appeared they did not share the same developments of change in opinions as the others. They experienced no change in the level of understanding despite the change in the duration of the clips. This might simply indicate that exposure to synthetic speech did not affect them in the same way.

Question 3: Was this form of speech tiring to listen to?

Participants:		3			7			9			10			11			12			14			17		
Gender:		M			M			M			M			M			M			M			M		
Duration:		L	M	S	L	M	S	L	M	S	L	M	S	L	M	S	L	M	S	L	M	S	L	M	S
Audio Quality:																									
tiring	levels 1-5	2	2	4	4	2	4	2	2	3	2	4	4	1	1	1	4	3	5	1	4	4	2	2	3

Figure 4.7 - Extract of raw data from male participants of question 3 of the questionnaire

Order: L→M→S:

The majority of the participants indicated a clear decrease in tiredness throughout the sequence of testing. While listening to long clips, 2 out of 4 responded that the synthetic speech was “tiring” and the other two responded with “very tiring”. When presented with clips of medium duration, opinions didn’t change much, except from one participant, who found a positive change going from “very tiring” to “not very tiring”. Moving along to the short clips in the end, the majority found these clips to be the *least* tiring to listen to and rated the level of tiredness as “not very tiring” and “average”.

Order: S→M→L:

Most of the participants (3 out of 4) showed no real sign of finding the synthetic speech tiring while listening to their first set of short clips. After listening to both medium and long clips however, the level of tiredness increased. Although some of them found the medium clip to be the slightly more tiring, and the short and long clips better, the overall trend is that the participants were more tired when listening to the long clips than when starting off with the short clips.

L→M→S and S→M→L:

As many as 87.5 % of the participants (7 out of 8) found the short clips to be the least tiring to listen to of all durations. A reason for this may simply be because the clips are so short that the quality of the synthetic speech does not bother/affect the testers as much as a medium or long clip might. In addition to finding the shortest clips least tiring, 6 out of 8 participants also found that the long clips had the highest level of tiredness. All in all there seems to be a trend amongst the participants in this group that the shorter the clips are, the less tiring they appeared to be.

Question 4: Could you listen to this form of speech for a longer period of time without wanting to stop?

Participants:			3		7			9			10			11			12			14			17		
Gender:			M		M			M			M			M			M			M			M		
Duration:		L	M	S	L	M	S	L	M	S	L	M	S	L	M	S	L	M	S	L	M	S	L	M	S
Audio Quality:																									
willingness to listen	leves 1-4	1	1	1	3	2	2	1	1	1	2	2	2	1	1	1	2	3	4	1	3	4	1	1	2

Figure 4.8 - Extract of raw data from male participants of question 4 of the questionnaire

Order: L→M→S:

After listening to the long clips, it was rather clear from all 4 of the participants that there was "no way" they would be willing to listen to this form of speech for a longer period of time. When presented with the medium and short clips, 3 out of 4 had the same opinion. The last one, however, showed signs of increasing willingness starting with "no way" for long clips, and

then moving on to “maybe” for medium clips and finally choosing “yes” for short clips. This participant explained that although the quality of the speech could be better, this did not affect his willingness to listen to longer clips. It was rather because of the detailed contents about historical monuments that he simply didn’t want to listen to very long clips.

Upon asking the participants why they were reluctant to listen to this form of speech for a longer period of time, they mentioned that the audio clips were hard to understand at times. They all agreed that the main reason for this was the speed of the speech, as well as the poor pronunciation of some words. Improving these aspects would most likely increase their willingness to listen. As for how long they would be able to listen, someone suggested a limit of the clips of no more than 3-4 minutes, but that the quality of the speech played a vital role for this to be possible.

Order: S→M→L:

The results from the participants that listened to audio clips in this order were more mixed than the results from the previous section. In the same manner as explained in the section above, the majority of the participants (3 out of 4) in this sequence were also rather reluctant to listen to this form of speech for a longer period of time. There are, however, mixed results with regards to responses for the different durations of the clips. Some got less and less willing to listen the further into the sequence they got, while others consistently answered “not likely” for all the clips, regardless of duration. There was also one participant that showed signs of increase in willingness to listen during the sequence. Although answering “not likely” after listening to short and medium clips, he seemed to grow more accustomed to this “*strange type of speech*” and answered “maybe” when listening to the long clip in the end.

L→M→S and S→M→L:

Comparing all the results of the male participants, it becomes clear that the majority of them were rather determined that listening to this type of speech for a longer period of time was out of the question, or at least not very likely. Listening to a clip for over one minute seemed more than enough time to get an idea of what was being described, and most participants started getting somewhat restless after listening for longer than this.

When talking with the participants after they had answered the questionnaire, most of the participants agreed that it was not necessarily the *duration* of the clip that was the problem, but rather the use of the "unnatural" *synthetic speech* for that duration that affected their opinions. With improvements to the current quality of the speech, they agreed that they might be more willing to listen to clips of longer duration, although not much longer. A couple of the participants, however, did not feel that the quality of the synthetic speech was much of a problem. In contrast, they found the clips rather interesting to listen to. They both agreed that if the speech was improved it would make listening easier, but that it would not necessarily make them more willing to listen to longer clips. To them one minute seemed like more than enough time to listen to someone describing a monument.

Before continuing on with the mobile-based testing, a brief overview of the summary data from the male and female participants from the computer-based testing will be given, represented in figures 4.9 and 4.10.

Summary Data - Female - Computer-Based Test					
		Avg	Min	Max	Mode
Audio Quality:					
choppiness	levels 1-7	3,19	1	6	3
understandable	levels 1-5	3,48	2	5	4
tiring	levels 1-5	3,37	2	5	4
willingness to listen	leves 1-4	2,7	2	4	2

Figure 4.9 – Female Summary Data

Summary Data - Male - Computer-Based Test					
		Avg	Min	Max	Mode
Audio Quality:					
choppiness	levels 1-7	2,65	1	6	3
understandable	levels 1-5	2,75	2	4	2
tiring	levels 1-5	2,75	1	5	2
willingness to listen	leves 1-4	1,79	1	4	1

Figure 4.10 – Male Summary Data

A closer look at these values reveals that the female results show signs of higher acceptance, or at least more positive results, compared to the male results. A possible reason for this may be derived on a common assumption that men tend to be more aware of technical/computer based issues than women. The problem with this assumption, however, is the fact that more and more women are getting involved with these types of issues, which may therefore rule out the assumption that "men are more aware of "technical/computer based issues"".

Another reason may be derived from yet another common assumption that women are more verbal and outgoing then men. From this, women may have met more dialects then men, and therefore be more positive when introduced to now “dialects”. A more likely reason for these results is based on the fact that there were simply not enough participants to generate a representative result set for basing any concrete conclusions. However, the low number of participants does give an idea of possible trends as to how users may react and respond to synthetic speech.

While conducting the actual testing, the age and educational level of all the participants were recorded. After analyzing the data with regards to these factors it was concluded that these factors had little, or no, impact as to how participants answered the questionnaire. This was confirmed by the fact that participants of equal age and/or educational level answered in some cases equally and in other cases not. Because there was no consistent pattern to their answers, this may be an indication that the age and educational level has relatively little (or nothing) to say as to how users may react when confronted with synthetic speech.

The *indications* that have been detected in this section will be presented in a more systematic matter in the chapter summary (section 4.4). Before doing this, the following section will present the results from the mobile-based testing.

4.2 MOBILE-BASED TESTING

When the mobile-based testing was complete a total of 8 participants were recorded. Out of these, there were 4 male and 4 female participants. The results from the mobile-based testing can be found in appendix F.

The organization of this section is the same as for the computer-based testing; the female results will be presented in section 4.2.1, followed by the male results in section 4.2.2.

4.2.1 RESULTS FROM FEMALE PARTICIPANTS

Out of the 4 female participants that were used during the mobile-based testing, 2 of them listened to audio clips in the order long-medium-short and the remaining 2 listened in the order short-medium-long.

Question 1) Was the speech smooth or choppy?

Participants:		2_3			2_4			2_7			2_8		
Gender:			F			F			F			F	
Duration:		L	M	S	L	M	S	L	M	S	L	M	S
Audio Quality:													
choppiness	levels 1-7	1	4	4	5	5	3	4	5	6	3	5	4

Figure 4.11 - Extract of raw data from female participants of question 1 of the questionnaire

Order: L→M→S:

Out of the two participants that listened to clips in this order there is a clear sign that the more they listen to the synthetic speech, the less choppy they consider it to be. Although there was a difference in their responses to long clips, the development from long to short clips is rather equal. One of them started out by finding the synthetic speech in long clips to be very choppy, whereas the other one started with a more positive response than this finding the clips to be of average choppiness. From here on and out, the responses were somewhat similar for both women. They considered the clip of medium duration to be smoother than the long ones, and the smallest clips to be even better. This might indicate that shorter clips are being accepted to a larger extent due precisely to their short durations. However, another explanation was also given from the participants and this is in line with the results presented earlier; the more they listened to this kind of speech, the less choppy it became.

Order: S→M→L:

One out of the two participants showed similar signs as the women above; rating the speech as less choppy the more she listened to it. This could mean that it is not necessarily the *duration* of the clips that affects the participants' response, but rather the *exposure* to the synthetic speech. In contrast, the second of the two participants started off rating the short clip as "average", then finding medium clips to be even less choppy, but when listening to clips of the longest duration there was a sudden increase in the rating of the choppiness. It was later explained after testing that the synthetic speech, although easier to understand

after listened to for a while, seemed to be somewhat choppier due to the audio clips having many foreign words and causing slight confusion.

L→M→S and S→M→L:

Evidences from the results indicate that combined, despite the order of the audio clips, most of the participants (3 out of 4) found that the more they listened to the clips the less choppy they seemed to be. It is quite interesting that the *order* in which the clips were played had very little influence on the participants' rating of the choppiness. The majority of the participants showed little, or no, difference in responding to short and medium clips, however between medium and long clips the level of choppiness in some cases rose again. In fact, upon discussing the decrease in choppiness, participants were surprised to hear that there was no difference between the ways the various clips were produced.

Question 2) Was the speech difficult to understand (was it unclear what was being said)?

Participants:		**2_3**			**2_4**			**2_7**			**2_8**		
Gender:			**F**			**F**			**F**			**F**	
Duration:		**L**	**M**	**S**	**L**	**M**	**S**	**L**	**M**	**S**	**L**	**M**	**S**
Audio Quality:													
understandable	levels 1-5	2	3	4	4	4	3	4	4	5	3	4	2

Figure 4.12 - Extract of raw data from female participants of question 2 of the questionnaire

Order: L→M→S:

Although starting off with different opinions, both of the participants that listened to this order of clips showed equal signs of an increase in understandability of the synthetic speech throughout the sequence. As mentioned in the results from question 1, this may be caused simply by the fact that after listening to synthetic speech for some time people get more accustomed to the voice and the way things are said.

Order: S→M→L:

In an almost similar manner as those who started off with the long clips, here too are signs that *experience* with listening to synthetic speech adds to a better understanding of what is being said. One of the participants rated the small clip as "average", and then showed an

increase by rating the speech as easy/clear for both the medium and the long clips. The other one also found the speech easier to understand after listening to the short clips, although a slight decrease was experienced between medium and long clips.

L→M→S and S→M→L:

Looking at all the results for this question combined, all of the participants (4 out of 4) responded with an *increase* in the rating of understandability towards the synthetic speech the further into the sequence they got. In most cases, this result was not affected by the order in which the clips were presented. There was, however, the issue of one participant experiencing a slight decrease when moving from the medium to the long clips, but seeing as the rating of the long clips was more positive than the rating of the short clips in the beginning of the sequence, there was still an overall increase in the understandability.

Question 3) Was the speech tiring to listen to?

Participants:		**2_3**			**2_4**			**2_7**			**2_8**		
Gender:			**F**			**F**			**F**			**F**	
Duration:		**L**	**M**	**S**	**L**	**M**	**S**	**L**	**M**	**S**	**L**	**M**	**S**
Audio Quality:													
tiring	levels 1-5	3	4	5	4	4	4	2	4	5	3	3	4

Figure 4.13 - Extract of raw data from female participants of question 3 of the questionnaire

Order: L→M→S:

Results from both participants were almost identical here. Upon listening to long clips, responses varied somewhat between “tiring” and “average”. However, after listening to clips of medium duration both participants found a decrease in tiredness responding with “not very tiring”. This positive development continued between medium and short clips as well, resulting with both participants finding the short clips to be “not tiring at all”.

S→M→L:

Results here varied slightly compared to those above. One of the two participants did not seem to be affected in any way as to how tiring the audio clips were. She was consistent throughout the whole sequence, finding all of the clips to be “not very tiring”. She simply

wasn't affected in any way with regards to the duration or quality of the speech. As for the second participant, her results indicated a marginal increase in tiredness. It was later brought to light that it was not the *duration* of the clips that made them slightly more tiring, but rather the sometimes poor *quality* of the speech.

L→M→S and S→M→L:

Combining the results above, it can be seen that the majority of the participants (3 out of 4) had relatively positive responses to the level of tiredness after listening to synthetic speech. The participants showed either a constant decrease in tiredness during the process, or remained consistent from the very beginning and did not get tired of the speech.

Question 4: Could you listen to this form of speech for a longer period of time without wanting to stop?

Participants:		2_3			2_4			2_7			2_8		
Gender:			F			F			F			F	
Duration:		L	M	S	L	M	S	L	M	S	L	M	S
Audio Quality:													
willingness to listen	leves 1-4	3	4	4	2	3	3	3	4	4	2	4	4

Figure 4.14 - Extract of raw data from female participants of question 4 of the questionnaire

Order: L→M→S:

Both the participants here found that the shorter the duration of the clips, the more willing they would be to listen to this form of speech for a longer period of time. In other words, the results show indications that the shorter the duration, the better. The long clips were not looked upon as being too long; however, the participants commented that making the clips much longer than they already were could possibly result in much more negative responses.

Order: S→M→L:

Similar to the above, both of the participants found that the longer the duration of the clips, the less willing they would be to listen to this form of speech for a longer period of time. One main difference from above, is that participants in this group were less willing to listen to clips

of long duration. Upon further discussion it was mentioned that an improvement of the quality of synthetic speech might lead to users being more willing to listen.

L→M→S and S→M→L:

Looking at the results from all 4 participants, they all agreed to the fact that the shorter the clips, the more willing they would be to consider listening to more of this speech. It was also unanimous that the clips of medium and short duration were of a suitable duration, while the longer clips were too long.

4.2.2 RESULTS FROM MALE PARTICIPANTS

As with the female testers above, there were also 4 male testers during this form of testing. Out of these, 2 listened to clips in the order long-medium-short and the remaining 2 listened in the order short-medium-long.

Question 1: Was the speech smooth or choppy?

Participants:		2_1			2_2			2_5			2_6		
Gender:			M			M			M			M	
Duration:		L	M	S	L	M	S	L	M	S	L	M	S
Audio Quality:													
choppiness	levels 1-7	1	2	3	3	5	1	3	4	5	3	4	2

Figure 4.15 - Extract of raw data from male participants of question 1 of the questionnaire

Order: L→M→S:

From the results of this order there seems to be a pattern that the shorter the clips became, the smoother the participants seemed to perceive this quality aspect of synthetic speech. This is confirmed by 2 out of 2 participants, experiencing a constant increase in smoothness the further into the sequence they got.

Order: S→M→L:

The overall results from this order also indicate an increase in smoothness; the more the participants listened to the synthetic speech, the more positive the responses became. One of the participants, however, experienced a greater jump (by two points) in the choppiness

when moving from the medium to the long clips. Despite this, the participant still had an overall improvement in the response from start to end. The "jump" was later explained as the long clips containing what seemed to be more foreign words than the other clips, resulting in confusion and the participant "falling off track" some times.

L→M→S and S→M→L:

Combined, all of the participants (4 out of 4) found an *increase* in the smoothness of the audio clips during the testing process, with the exception of the one who found long clips to be choppier than medium clips. Given the order of the clips, it appears that it is not whether or not participants listen to short, medium or long clips first that plays a role in their responses, but rather the degree of *exposure*. After participants had listened to synthetic speech for some time they seemed to become more aware of how the voice behaved, hence getting a better understanding of what to expect.

Question 2: Was the speech difficult to understand (was it unclear what was being said)?

Participants:		2_1			2_2			2_5			2_6		
Gender:			M			M			M			M	
Duration:		L	M	S	L	M	S	L	M	S	L	M	S
Audio Quality:													
understandable	levels 1-5	2	3	3	3	4	3	3	4	4	3	3	3

Figure 4.16 - Extract of raw data from male participants of question 2 of the questionnaire

Order: L→M→S:

As for the understandability of what was being said, both participants found that the shorter the clips were, the easier it became to understand what was being said. They both agreed that the contents of the short clips also seemed clearer, something that might have affected their response.

Order: S→M→L:

1 out of the 2 participants rated the short clips as "average" and then responded more positive ("easy") to the medium clips. However, when responding to the long clips, the response fell back down to "average". This could be due to the fact that longer clips have a

higher chance of containing words that are poorly converted by the TTS engine, or it could even be the use of more foreign words than in the other clips. The second participant simply found all clips to be equally easy to understand, indicating no problems with the contents of the clips.

L→M→S and S→M→L:

Out of all the 4 participants, the majority (3 out of 4) indicates that over time, and somewhat despite the order in which clips were played, the clips became easier to understand, or they were at least equally understandable throughout the sequence. As stated earlier, there seems to be overall indications that *exposure* to, and experience with, synthetic speech seems to increase participants understanding of what is being said in the clips.

Question 3: Was the speech tiring to listen to?

Participants:		2_1			2_2			2_5			2_6		
Gender:			M			M			M			M	
Duration:		L	M	S	L	M	S	L	M	S	L	M	S
Audio Quality:													
tiring	levels 1-5	2	3	3	4	4	3	2	4	4	4	4	3

Figure 4.17 - Extract of raw data from male participants of question 3 of the questionnaire

Order: L→M→S:

Both participants that listened to this order of clips found that the longer the clips were, the more tiring they seemed to be. In contrast, clips of short and medium durations were rated as less tiring.

Order: S→M→L:

As opposed to the participants that listened to clips from long to short, both participants here indicated a *decrease* in tiredness the longer the clips became. They were both unanimous with how tiring they experienced the clips to be. As indicated several times, from many of the participants, this can be seen as a possible trend that the more they listen and get accustomed to the new form of speech, the less tiring the clips seem to get while listening.

L→M→S and S→M→L:

Comparing all the results from question 3 only strengthens the trend mentioned several times earlier. All the participants find that the speech becomes less and less tiring the more they listen to it. In the first case (L→M→S), this means that the short clips are the least tiring, while in the second case (S→M→L) it means that the long clips are the least tiring. In both cases, it is the clips towards the *end* of the sequence that are rated as the best ones. This makes perfect sense seen in the light of the earlier mentioned trend that more experience with the speech, or more *exposure*, improves the responses from the participants.

Question 4: Could you listen to this form of speech for a longer period of time without wanting to stop?

Participants:		2_1			2_2			2_5			2_6		
Gender:			M			M			M			M	
Duration:		L	M	S	L	M	S	L	M	S	L	M	S
Audio Quality:													
willingness to listen	leves 1-4	3	4	4	4	4	4	2	4	4	3	3	3

Figure 4.18 - Extract of raw data from male participants of question 4 of the questionnaire

Order: L→M→S:

As mentioned on question 4 in section 5.1.1, the focus in this question was mainly on the *duration* of the clips and not the quality of the speech. It was simply meant to investigate if participants found the defined durations acceptable or too long. Both participants that listened to clips in this order found the longest clips slightly tiring and were only "maybe" willing to listen to this form of speech for even longer. After listening to the short and medium clips, however, they responded "yes" to the question of whether they could listen to this for a longer period of time.

Order: S→M→L:

1 of the 2 participants answered "maybe" for all three sets of durations, while the other one was slightly more negative responding "not likely". As to what could be done to improve the

willingness to listen to longer clips, one of the participants suggested to simply slow down the *speed* of the speech. By slowing things down it would be easier to follow the voice and thereby, possibly, increase the willingness to listen. The second participant simply referred to improving the *quality* of the synthetic speech, like many other participants have suggested earlier.

L→M→S and S→M→L:

Looking at the responses from all 4 participants, there is no real decrease in the willingness to keep listening, regardless of the duration of the clip they listen to. However, there is an indication that the long clips were given the most negative responses. As mentioned above, the primary solutions brought to light were those of slowing down the speed of the speech as well as improving the quality of the speech. More light will be shed on these suggestions in chapter 6.

Before moving on to the combined summary data of *both* test sessions, a brief look at the summary data from the male and female participants from the mobile-based testing is presented in figures 4.19 and 4.20.

Summary Data Female Mobile-Based Test					
		Avg	Min	Max	Mode
Audio Quality:					
choppiness	levels 1-7	4,08	1	6	3
understandable	levels 1-5	3,5	2	5	4
tiring	levels 1-5	3,75	2	5	4
willingness to listen	leves 1-4	3,33	2	4	4

Figure 4.19 – Female Summary Data

Summary Data Male Mobile-Based Test					
		Avg	Min	Max	Mode
Audio Quality:					
choppiness	levels 1-7	3	1	5	3
understandable	levels 1-5	3,17	2	4	3
tiring	levels 1-5	3,33	2	4	4
willingness to listen	leves 1-4	3,5	2	4	4

Figure 4.20 – Male Summary Data

Taking a closer look at these values it can be seen that the female results once again show signs of higher acceptance, or at least more positive results, compared to those of the male results. This was also the case in the summary data figures presented for male and female

participants of the computer-based testing. A reason for these results may be the fact that there simply weren't many participants to generate a proper set of results to correctly represent the differences between genders. However, the fact that the results show higher acceptance by females in *both* test forms might not be a coincidence. It is possible that this actually indicates a general trend, and not necessarily a lack of participants.

4.3 COMBINED SUMMARY DATA FOR BOTH TEST SESSIONS

The main reason for using *two* forms of testing, computer-based and mobile-based, was to find out if the environment surrounding the participants affected their judgment towards answering the questions in the questionnaire. By *environment*, it is meant whether the participants were in front of a computer or walking around Bergen during testing.

Figures 4.21 and 4.22 below present the combined summary data from both test sessions.

Summary Data - Combined - Comp.-Based Test					
		Avg	Min	Max	Mode
Audio Quality:					
choppiness	levels 1-7	2,92	1	6	3
understandable	levels 1-5	3,14	2	5	4
tiring	levels 1-5	3,08	1	5	4
willingness to listen	leves 1-4	2,27	1	4	2

Figure 4.21 – Summary data of combined computer-based testing

Summary Data Combined Mobile-Based Test					
		Avg	Min	Max	Mode
Audio Quality:					
choppiness	levels 1-7	3,54	1	6	3
understandable	levels 1-5	3,32	2	5	3
tiring	levels 1-5	3,54	2	5	4
willingness to listen	leves 1-4	3,34	2	4	4

Figure 4.22 – Summary data of combined mobile-based testing

By looking at the figures, some may say that there appeared (in some cases) to be a large (significant) difference leading to assume that the environments did in fact have something to say towards their responses to questions 1-4 from the questionnaire. However, this may be debatable. After conducting a paired *t* test, utilizing a free online *t* test calculator[14], to

[14] The calculator used here, as well as some information as to what a t test is can be found at GraphPad Softwares' homepage (link found in bibliography).

calculate the *statistical* significance of the two sets of data, it was revealed[15] that the difference was in fact not quite (statistically) significant. In short, the t test is used to compare the means of two sets of data, in this case the values gathered from the mobile and computer-based testing, to examine whether or not there is a (statistical) significant difference in the two sets. In other words, statistically, there was no significant difference in the gathered data based on the environments they were collected in. Although results from the mobile-based testing are slightly higher than those of the computer-based testing, this may be explained due to the difference in the number of participants used in each test session. More exactly, the mobile-based testing used almost half the number of participants as the computer-based testing (respectively 8 participants versus 17 participants).

To the extent that these results can be generalized, one can say that it is not necessary to divide the participants into groups of different environments before conducting the testing. Because of the similarity (lack of significant difference) in averages that was found by doing this, an easier way of testing may be to only have one large group of participants instead of two smaller ones. In other words, the indications from questions 1-4 presented in this chapter would, to a large extent, have been the same if all of the 25 participants were gathered in one group. This might be a useful insight when planning to conduct testing in the future.

4.4 CHAPTER SUMMARY

The analysis of the two test sessions, presented in this chapter, included very detailed information about the results gathered, as well as several different trends and indications. In order to organize some of this information, it will now be given a short summary of the most important indications. Because the female and male participants in both the computer-based and the mobile-based testing responded in a very similar way, the indications that arose from their responses were close to identical. Therefore the summary will not separate the indications for female and male or the indications for the two different test sessions, but look at the *summarized indications for all of the participants*.

[15] Screenshot of the results gathered from the t test can be found in appendix G.

First and foremost, the main indication that needs to be remembered is the fact that *exposure* to synthetic speech appears to increase the level of acceptance of the speech. In other words, the more the participants listened to this form of speech, the more accustomed they got to it and the more positive their responses became. They perceive the speech as less choppy, less tiring and more understandable the further into the testing sequence they get. It can be derived from this that the *duration* of the clips does not matter much, at least in most of the cases.

More often than not, the *order* of the clips didn't affect the participants in any significant way. When listening to clips in the order long-medium-short many of them found the short clips to be the easiest to understand, while when listening to the order short-medium-long the long clips were perceived to be the easiest. As can be seen, it is the clips towards the *end* of the sequence that receive the most positive response. This can be explained based on the fact that towards the end, the *exposure* to the speech is at its highest. It is therefore not the order in which the clips are presented that matters, but how long they have been listening to the synthetic speech and how accustomed they have become to it.

Continuing, it should be mentioned that neither age nor the level of education of the participants played a vital role in the test sessions. This was discovered because participants of the *same* age and education level answered somewhat *different* in many cases. In addition, participants of *different* ages and education levels responded somewhat *similarly* on many of the questions. Because of this, it did not emerge any form of evidence that these two factors affected the results.

Finally, two additional interesting insights came to light in this chapter. One was that the female participants showed a higher level of acceptance of the synthetic speech, or at least more positive responses, than the male participants. This was the case in both the computer-based and the mobile-based testing. The other insight was the fact that dividing the participants into these two groups of testing in order for them to be surrounded by different environments, had no significant effect on their responses, as discovered through the *t* test.

5 EVALUATION AND CONCLUSION

The main focus of this thesis has been to investigate the acceptance of synthetic speech in order to uncover possible solutions as to how the *acceptance* of this form of speech can be increased. This intention was also specified in the research question.

The last two questions in the questionnaire (question 5 and 6) were asked to find out the participants' suggestions and comments as to how the synthetic speech could be improved, and thereby also how the participants' acceptance of the speech would increase. The questions were as follows:

5. How do you think this form of speech could be improved?
6. Any other comments?

The responses from these questions, as well as the post experiment interviews, were vital to be able to answer the research question, and will therefore be presented in detail in section 5.1. Section 5.2 evaluates the hypothesis presented in the beginning of the thesis, while section 5.3 will present the final conclusions. At the end of the chapter, section 5.4 gives an overview over issues possibly worth looking into for future research into the subject of synthetic speech.

5.1 EVALUATION OF TESTING

Given the fact that there were relatively few participants used during the course of the two test sessions, a clear and "representative" conclusion can be difficult, if not impossible, to produce. However, conclusions may be derived as *indications* based on the suggestions for improvement that the majority of the participants agreed on. In other words, if many of the participants come up with the exact same suggestions, this would give some clear indications as to what aspects of the synthetic speech that need to be changed.

Because the female and male participants came with very similar suggestions for improvement of the synthetic speech, it was not considered necessary to keep their responses separated. Instead, all the suggestions that the *majority* of the participants agreed

on have been added together and will be presented in section 5.1.1. For a complete overview of the separate responses from the females and males, please refer to appendix F. Also, the following sections will only present figures of *summary* data from the different aspects of the testing. For a more detailed overview of the *raw* data, see appendix E.

5.1.1 COMBINED SUGGESTIONS FOR IMPROVEMENT

Figure 5.1 and 5.2 show the combined suggestions from females and males, from respectively the computer-based and the mobile-based testing, as to how synthetic speech can be improved. In other words, it shows their responses to question 5; "How do you think this form of speech could be improved?".

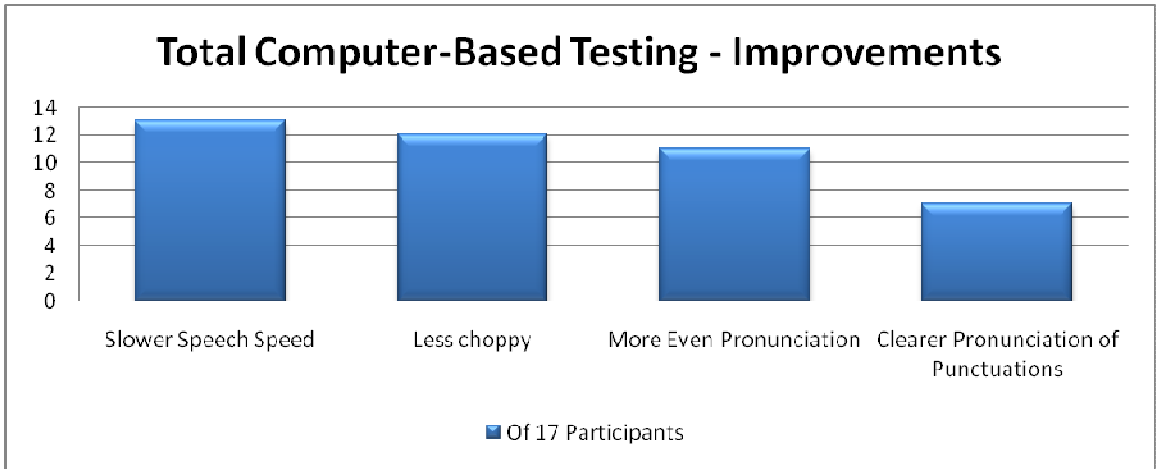

Figure 5.1 – Combined suggestions for improvement from computer-based testing

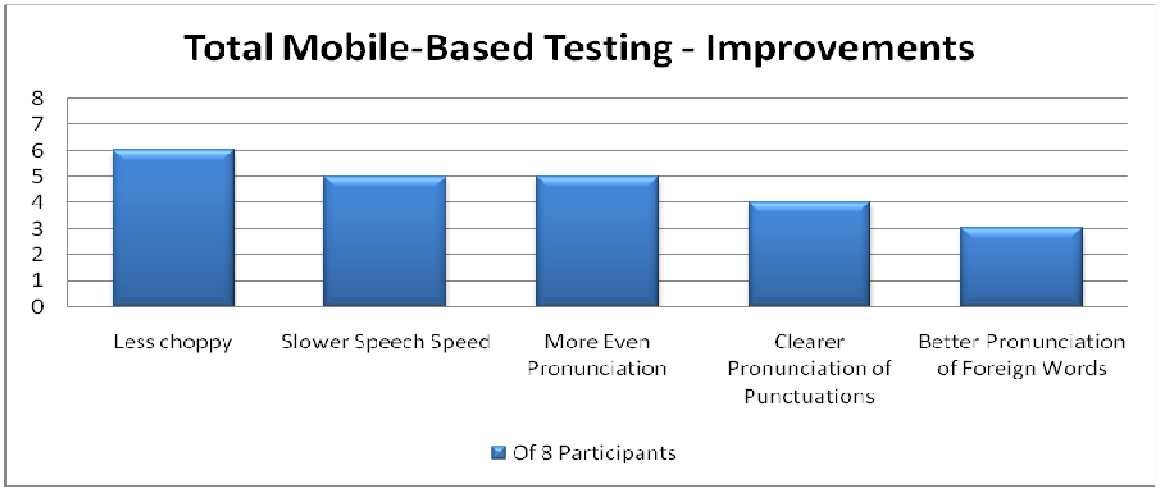

Figure 5.2 – Combined suggestions for improvement from mobile-based testing

The two graphs were made by taking all the suggestions mentioned in the individual graphs for female and male graphs and combining them[16]. As can be seen, the suggestions from both groups of participants are very equal. The only difference is that in the mobile-based testing, an additional suggestion is mentioned; better pronunciations of foreign words. There were, however, only three of the participants that mentioned this. Because of the similarities of the two graphs, it seems appropriate to combine them to get a complete overview of which suggestions that have been most frequently mentioned by all the 25 participants. Figure 5.3 shows the *total* frequency of the suggested improvements.

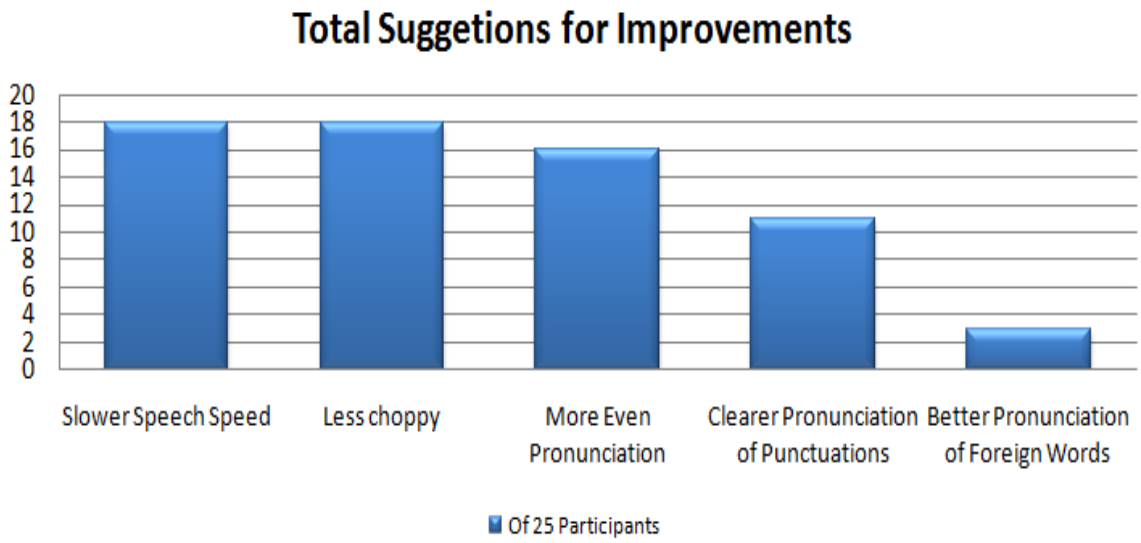

Figure 5.3 – Total summary of suggestions for improvements

Keep in mind that this graph represents the results from all 25 participants from the testing phase of this thesis. If a large amount of these participants agree on that a certain aspect of the synthetic speech system needs improvement, this is worth noticing. In other words, it isn't possible to draw any conclusions based on what a *few* participants think, but whatever the *majority* thinks should be paid more attention to.

Overall, as many as 72% of the participants (18 out of 25) agreed that the *speed of the speech* could be improved. By changing the speed, they meant slowing it down to make it easier to

[16] All of these graphs can be found in appendix F.

follow what was being said. To actually execute this change is not much of a problem, because the majority of TTS engines available today have a functionality that makes it possible to adjust the speed when converting a text. An example of this can be found back in section 3.1.1, where figure 3.1 presented how to use the TTS engine used in this thesis. It was explained that one could simply adjust a scroll bar until the desired speed was reached. Before actually converting the text, one could test the different levels of speed to see which level sounded the best, and then continue with the conversion. In the case of this thesis, the speed of the synthetic speech was set to the default (zero) value of the TTS engine. Therefore, following the suggestions of the participants would involve setting the speed value in the TTS engine down 1 or 2 levels, to -1 or -2. As to which of these values would be best, or most accepted, is unknown and would have to be tested, possibly by performing the proposed testing in this thesis over again, but simply concentrating on the speed aspect of the speech.

Moving on, 72 % of the participants also agreed that the *level of choppiness* in the speech should be improved. Reducing the choppiness would involve changing how words are strung together by the TTS engine. Several factors come into play here, such as improving the pronunciation of words[17]. In order to do this, normalization of the text is first required, which mainly involves expanding numbers, special symbols, acronyms and abbreviations. Once this is done, improving word pronunciation can begin. Using better algorithms and rules for pronunciation involving, for example, *phoneme mapping*, seem to be one of the major factors in reducing the choppiness of synthetic speech. However, such improvements might require using a commercial TTS engine which would have to be purchased in order to be used. An example of such a commercial engine is *Loquendo TTS* (Baggoa et al., 2006).

The next aspect of improvement, that a total of 64 % of the participants agreed on (16 out of 25), is that of *more even pronunciation*. In other words, that the speech should sound more natural than it does at this stage. Upon explaining this, the participants mentioned that sometimes the text was pronounced as a question when it was simply a statement. Some of them also gave the example that the synthetic voice used the same pitch for a specific word regardless of the context the word was said in, making the pronunciation seem more uneven.

[17] Loquendo's homepage

Moving on, almost half of the participants (44 %) agreed to the fact that *pronunciation and punctuation should be clearer*. In particular, they mentioned that the improvement of punctuations, like ".", "!" and "?", should be emphasized. In several cases they found it very difficult to hear the end of a sentence and the start of a new one, resulting in an increased level of tiredness after listening for only a short period of time.

Finally, a few of the participants, 12 % in total, suggested that better *pronunciation of foreign words* could improve the quality of the speech. The reason that so few commented on this, may be due to the specifications of the TTS engine used in this thesis. Because the TTS engine is based on the English language it was necessary to explain to the participants that this factor was not to be considered throughout testing. This was furthered illustrated by an example; to imagine an English-speaking person trying to pronounce Norwegian names and places as they would be said in Norwegian. Not very promising some might say.

It is possible to solve this "multi-language" problem by using a TTS engine that can mix two (or more) different languages and recognize which language every word in the text belongs to. A good example of this is the commercial TTS engine *Loquendo TTS*, mentioned above, that presents the feature of Mixed-Language Support. Through the use of simple control tags added to an input text, it is possible to change the language of the "speaker" while one is writing in order to improve the quality of how foreign words are pronounced. The language can be changed back to its original language at any time in the same way, ensuring a higher quality of the speech.

Unfortunately, the Mixed-Language Support in this TTS engine does not include the Norwegian language. It would therefore not have been able to improve the pronunciations of Norwegian words in an English text, as is the case of this thesis. Although the engine includes the Swedish and Danish language, these would still not be ideal substitutes for the Norwegian language. In addition, this engine is a commercial product and would have to be purchased in order to actually be used.

To sum up this section, the following list contains the suggestions for improvement that were mentioned by the participants.

- Slowing down the speed of the synthetic speech (72 % mentioned this)
- Make the speech less choppy (72 %)
- Make the pronunciation of the words more even (64 %)
- Make the pronunciation, and in particular the punctuations, more clear (44 %)
- Improve the pronunciation of foreign word (12 %)

In order to learn even more from the participants, and possibly get more detailed information from them, they were given a last opportunity to comment on the improvements of synthetic speech or any other aspects of the test sessions. These comments will now be presented in section 5.1.2.

5.1.2 OTHER COMMENTS FROM THE PARTICIPANTS

The complete list of responses to the last question, question 6: “Any other comments?”, was as follows:

- The speech is clear at times, but can jump to very choppy in the same sentence
- Foreign names and places were pronounced very unclearly
- Fell off track right after a foreign word was presented
- It seemed like the speed was changing from time to time
- The speed of the clips was at times too fast
- *Duration* wasn’t the problem with listening to longer clips, but rather the poor *quality* of the speech
- Over time the synthetic speech became easier to listen to and more understandable
- The speaker didn’t always pick up on punctuations
- The content of clips might affect willingness to listen to longer clips. For example, uninteresting facts about historical monuments can make it more tiring to listen in the long run.
- More willing to listen to longer clips if the quality of the speech is improved

Almost all of these comments occurred in *both* the computer-based and the mobile-based test sessions. Because of this, the responses from the two sessions were close to identical and they will therefore be combined into one figure. Figure 5.4 presents the final comments that were made the most *frequently* by the in total 25 participants.

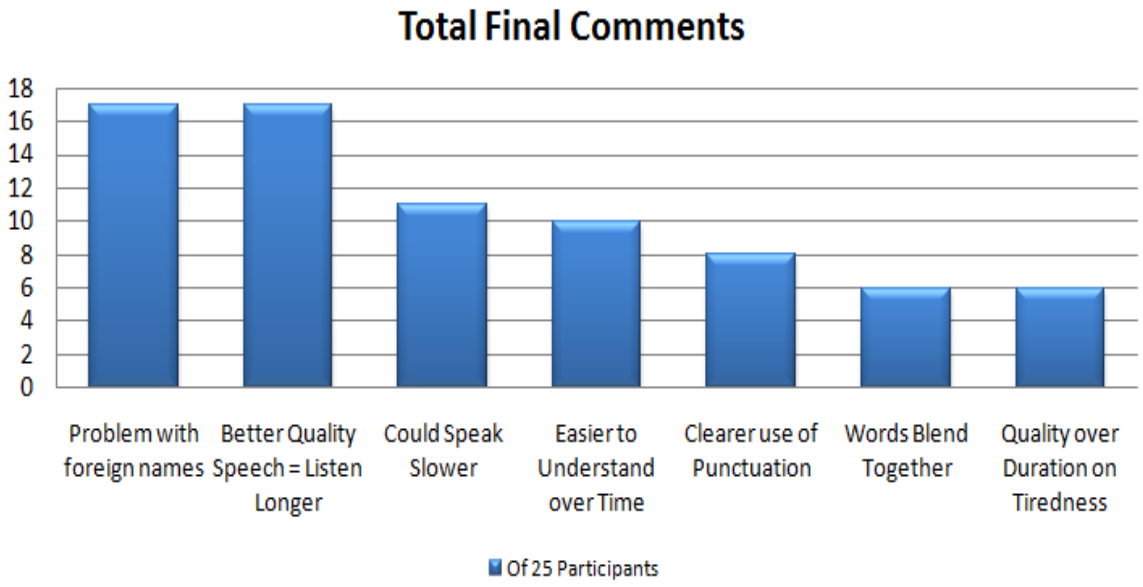

Figure 5.4 – Summary of final comments

Like the figure illustrates, there are two aspects of synthetic speech that are mentioned far more frequently than others. One aspect is the problem with *foreign words* in the clips, while the other one is the fact that participants will become more willing to listen to this form of speech if the *quality of the speech* improves. In fact, a total of 68 % of the participants (17 out of 25) agreed on both of these two aspects.

Regarding the pronunciation of *foreign words*, it was explained towards the end of section 5.1.1 that although there do exist ways of solving this problem for some languages, there are per today no engines that can solve the problem for the Norwegian language. Another solution could be to simply write foreign words the way they *sound*, however this would

defeat the purpose of an automatic converting engine as users would first have to hear the texts and alter them before converting them.

Moving on to the *quality of the synthetic speech*, this comment did not reveal anything new compared to what has already been mentioned about this aspect earlier. It simply means that the participants would find this form of speech less tiring, and more understandable, if there were made some changes. The participants' suggestions as to what these changes could be were presented in section 5.1.1, and the final comments made here only *emphasize* the importance of actually going through with these changes.

Finally, there were also quite a few comments regarding *punctuations* and the *speed of the speech*. Because the TTS engine sometimes does a poor job in considering punctuations in the text, and the speed of the speech seems to change rather often, the participants found that this resulted in *words blending together*. All of these comments are in line with the participants' suggestions to improvement presented in the previous section.

5.2 HYPOTHESIS EVALUATION

When the work on this thesis first started, a hypothesis was given in order to generate a basis for how synthetic speech could be improved. The hypothesis was as followed:

> *HYP: "Shorter audio clips make synthesized speech easier to listen to and thereby more accepted."*

Early on in the analysis of the collected user feedback, indications of the falsification of this hypothesis were already dawning. As the results of the analysis in chapter 5 indicated, simply reducing the *duration* of audio clips was not enough to increase user acceptance of synthetic speech. The participants did not always respond in a positive way when presented with short clips, and at times long clips were found to have more positive responses than short ones. The test results rather indicated that *exposure* to the speech to a large extent improved the participants' responses; the more they listened to it, the less choppy and tiring and the more understandable they perceived it. In other words, the duration of the clips did not matter much towards affecting the participants compared to the exposure to the speech itself.

On this ground, it is therefore possible to conclude that shorter segmented clips did *not* necessarily make it easier to listen to synthetic speech and thereby make it more accepted. Put differently, the hypothesis is not correct. Due to limitations on the number of participants, one should be careful to conclude with 100 % confidence that the hypothesis has been proven false, although from the results gathered there is at least a very good reason to believe this.

5.3 CONCLUSION

At this point, a conclusion may be given with regards to the formulated research question from section 1.3.1:

> *RQ: "How can the acceptance of synthetic speech be improved?"*

Results from analyzing and evaluating the user feedback from testing show a clear indication as to how this research question may be answered. Based on this, it seems accurate to conclude that simply *exposing* possible users to synthetic speech in an attempt to accustom and prepare them for something other than a normal human voice, can *strongly increase* the user acceptance of synthetic speech. In addition, it would also be necessary to improve the various discussed *quality aspects* of the speech to insure an even larger increase in acceptance.

5.4 FUTURE RESEARCH

As it has been discussed in chapter 5, the number of participants used during testing was rather limited. In order to conduct a more representative evaluation of the effects of the various audio qualities that have been presented in this thesis, testing should be repeated in a larger scale. Using more participants for testing, as well as trying to broaden the age range, could insure a higher chance that all possible users of synthetic speech will be considered. However, even if testing in the future is performed on a smaller scale, say, with 25 participants like in this thesis, it can still produce valuable information. If there are conducted a large amount of tests like this, and they all show the same results, then together their implications would appear as more credible and might very well be valid.

As the audio files used in this thesis were generated by a text-to-speech engine available for free on the Internet, the issue of using a product that requires purchasing was brought to light. Assuming a product that needs to be purchased before being able to use/test it could give better results (better quality conversions) than that of a free product, it could be interesting to compare user responses of converted texts from the different TTS engines and investigate whether or not any significant differences exist.

Mentioned by several of the participants, the occurrence of foreign words in the descriptions played during testing added to the level of difficulty to follow along, and possibly understand, what was being said in the clips. By keeping a record of the number of foreign words in each audio clip, a possible topic to look further into here could be to analyze the effect on users' responses based on these numbers of foreign words.

As it was difficult to find any relevant articles as to how to classify the three duration categories (long, medium and short), it could be interesting to first find or create some sort of theoretic reasoning for how audio clips, in a given context, should be classified. From this comes the possibility of better categorizing audio clips of a collection, with a higher chance of knowing how users themselves define the duration of audio clips.

In relation to how testing was conducted in this thesis there are several issues deserving attention as to how they might need to be done differently in the future. Primarily the issues revolve around the questionnaire used for gathering the necessary information for analyzing and evaluating. Revising the questions asked, normalizing the number of choices to choose from, as well as possibly adding new questions covering any other important aspect of synthetic speech not already taken into consideration, could add to a higher quality set of responses amongst participants. In addition to this, reorganizing the converted audio clips in terms of their context and duration could be conducted. An example of this could be:

- short clips: name, year made, maker
- medium clips: , name, year made, maker, materials used, measurements
- long clips: all of the above plus history

In doing so, it may be possible to better avoid the chances of users getting tired of listening to clips based on their content. In knowing the type of content of a clip based on its duration, users may be presented with information in a manner more to their liking.

Finally, a design for a prototype will now be presented as a means for using synthetic speech in a common real world situation.

A system for retrieving information (images) on historical monuments in the city of Bergen, Norway, exists today. This system was developed by members of the CAIM research group, and is known as the MMIR2 (Multi Media Information Retrieval) prototype (Hellevang, 2008). The MMIR2 is a system for image retrieval performed with the use of a mobile telephone (Nokia phones with GPS functionality). By taking a picture with a mobile phone and sending it to the MMIR2 system, one can receive images similar to the one sent in to the system. A result set consisting of four images is presented to the user, where they can select one and one image and be given an enlarged version to be able to view the image with greater detail.

As this prototype focuses on the use of images and not audio files, some changes would have to be made in order to be able to use it. The changes would involve altering the structure of the database containing all the information on specific monuments. This would be necessary in order to be able to add audio clips to the database and retrieve them when queried for. In addition, one would have to make changes to a Java client which is to be installed on a mobile phone. The objective of enhancing MMIR2 would be to add new functionality to perform an additional search for an images' corresponding audio description when looking at the enlarged version of an image.

Additional changes to the MMIR2 prototype could involve the use of information, in the form of both text and audio, displayed on mobile telephones to evaluate what users may prefer to be presented with when searching for information on monuments.

Bibliography

Books and articles:

Eide, E. et al. (2003): “Recent Improvements to the IBM Trainable Speech Synthesis System”. Volume 1. IBM Thomas. J. Watson Research Center, New York.

Gong, L. & Lai, J. (2001): “Shall We Mix Synthetic Speech and Human Speech?: Impact on Users’ Performance, Perception, and Attitude”. Association for Computing Machinery (ACM), New York.

Hellevang, M. (2008): “MMIR2 – Mobile Multimedia Image Retrieval”. Department of Information Science and Media Studies, University of Bergen.

Kjeldskov, J. & Graham, C. (2003): “A Review of Mobile HCI Research Methods”. Volume 2795/2003, pp. 317-335. Department of Information Systems, University of Melbourne.

Klatt, Dennis. H. (1987): “Review of text-to-speech conversion for English”. The Journal of the Acoustical Society of America, Volume 82, Issue 3, pp.737-793.

Langøy, M. (2008): “BergenBy Database & Metadata Editor”. Department of Information Science and Media Studies, University of Bergen.

Müller, M. et al. (2005): “Audio matching via chroma-based statistical features”. University of London.

Nass, C. & Min Lee, K. (2000): “Does Computer Generated Speech Manifest Personality”. Association for Computing Machinery (ACM), New York.

Nass, C. et al. (2000): “Can Computer-Generated Speech Have Gender? An Experimental Test of Gender Stereotype”. Department of Communication, Stanford University.

Nordbotten, Joan. C. (2006): “ADM - Advanced Data Management - Now: Multimedia Information Retrieval Systems”. Department of Information Science and Media Studies, Universuty of Bergen.

Owen, C. B. & Fillia Makedon (1999): “Cross-modal information retrieval”. In: *Handbook of internet and Multimedia: Systems and Applications*, pp. 109-129. CPC Press, Florida.

Zhiwei, S. et al. (2008): ”IBM Voice Conversion Systems for 2007 TC-STAR Evaluation”. Volume 13, number 4. IBM China Research Lab, Beijing, IBM Thomas. J. Watson Research Center, New York.

Internet Sources:

Acapela-Group’s homepage: “How Does It Work?”. <http://www.acapela-group.com/how-does-text-to-speech-work.html> (07.07.2009)

Audio Description’s homepage: “Art and Audio Description” <http://www.audiodescription.com.au/index.php?option=com_content&view=article&id=5&Itemid=7 > (20.10.2008)

GraphPad Softwares’ homepage: “Quick Calcs Online Calculators for Scientists – t test calculator”. <http://www.graphpad.com/quickcalcs/ttest1.cfm> (25.09.09)

Loquendo’s homepage: “ACHIEVING PERFECT TTS INTELLIGIBILITY”. <http://www.loquendo.com/en/brochure/AVIOS06_Loq_TTS.pdf> (23.06.2009)

Loquendo’s homepage: “White Paper – SSML 1.0: an XML-based language to improve TTS rendering”. <http://www.loquendo.com/en/whitepapers/SSML.1.0.pdf > (23.06.2009)

NaturalSoft’s homepage: “Free download”. <http://www.naturalreaders.com/download.htm> (19.11.2008).

PC Magazine’s homepage: “Encyclopedia”. <http://www.pcmag.com/encyclopedia> (11.09.2008)

ReadPlease's homepage: "Download".
<http://www.readplease.com/english/downloads/#rp2003> (19.11.2008).

Sayvoice's homepage: "Sayvoice Text-to-Speech for Windows".
<http://sayvoice.com/downloads.php> (19.11.2008).

SpeakComputers' homepage: "Free Text to Speech Software".
<http://www.speakcomputers.com/Text-to-Speech.aspx> (19.11.2008).

Speech Technology's homepage: "How TTS Works: The Technology Behind Text".
<http://www.speechtechmag.com/Articles/Editorial/Feature/How-TTS-Works-The-Technology-Behind-Text-29522.aspx> (24.03.2009)

TanseonSystem's homepage: "Download our Software Free!"
<http://tanseon.com/products/voicemx.htm> (19.11.2008).

TESL-EJ's homepage: "On the Internet – Text-to-Speech Applications Used in EFL Contexts to Enhace Pronunciation". <http://tesl-ej.org/ej42/int.html.bu2> (05.02.2009).

APPENDIX

APPENDIX A – INFORMATION SHEET

START

INFORMATION SHEET:

Research Title: Acceptance of Synthetic Speech in Multi-Modal Information Retrieval

The Department of Information Science

University of Bergen

You are invited to participate in a research experiment conducted by Andrew Moores.
The following experiment is being conducted to try and learn the reactions of users to the use of synthetic (computer generated) speech when receiving descriptive information on historical structures and monuments in Bergen. The context of this thesis involves the use of mobile devices (mobile telephones, DPAs) for receiving and listening to automatically generated audio clips. Participants are not expected to have any pre-knowledge of the field of this thesis. However, it is required to understand English as the audio clips used in this experiment are all stored in English.
If you should volunteer to Participate in this experiment, you will be asked to do the following tasks:

- Listen to a collection of pre-selected audio clips using headphones
- Fill out a questionnaire about the audio clips you heard
- Discuss of the experiment and other questions with the researcher

It is possible to end participation at any time.
The experiment will take place one participant at a time with the conductor and should take no more than 30 minutes total with no follow-ups.
The only personal information needed from participants will be their gender, age and educational (bachelor/master) degree. To assure the anonymity and to prevent the identification of all participants, no names will be recorded or stored and all other information will be treated confidentially.
END

APPENDIX B – PARTICIPANT CONCENT FORM

START

PARTICIPANT CONSENT FORM:

Research Title: Acceptance of Synthetic Speech in Multi-Modal Information Retrieval

The Department of Information Science

University of Bergen

Participant's Number: _____

Gender: (circle your option) Male / Female

Age: 18-22☐ 23-27☐ 28-32☐ 33-37☐ 38-42☐ 43-47☐

Educational Degree: _____

Date: ___/___/___

1. I have read the Information Sheet for this study and have had details of the study explained to me.
2. My questions about the study have been answered to my satisfaction, and I understand that I may ask further questions at any time.
3. I also understand that I am free to withdraw from the study at any time, or to decline to answer any particular questions in the study.
4. I agree to provide information to the researchers under the conditions of confidentiality set out on the information sheet.
5. I wish to participate in this study under the conditions set out in the Information Sheet.

☐ I have read and agree to the above statements

Researcher's Name: _____

Researcher's Signature: _____

END

APPENDIX C – QUESTIONNAIRE

START

QUESTIONNAIRE: Clips set: (circle option) 1 1* 2 2* 3 3* 4 4* 5 5*

Length: (circle option) Long Medium Short

Research Title: Acceptance of Synthetic Speech in Multi-Modal Information Retrieval

The Department of Information Science

University of Bergen

Please listen to the pre-selected collection of audio clips. When you are finished listening please answer the following questions as honestly as possible. If the given check-box options do not fit your opinion, or you should have any other comments, please add this in the "comment" section.

1. Was the speech smooth or choppy?

- ☐ Very choppy
- ☐ Choppy
- ☐ Slightly choppy
- ☐ Average
- ☐ Slightly smooth
- ☐ Smooth
- ☐ Very smooth

Comment:

2. Was the speech difficult to understand (was it clear/unclear what was being said)?

- ☐ Very difficult / Very unclear
- ☐ Difficult / unclear
- ☐ Average
- ☐ Easy / clear
- ☐ Very easy / Very clear

Comment:

3 Was this form of speech tiring to listen to?

- ☐ Very tiring
- ☐ Tiring
- ☐ Average
- ☐ Not very tiring
- ☐ Not at all tiring

Comment:

4 Could you listen to this form of speech for a longer period of time without wanting to stop?

- ☐ No way
- ☐ Not likely
- ☐ Maybe
- ☐ Yes

Comment:

5 How do you think this form of speech could be improved?

Comment:

6 Any other comments:

END

APPENDIX D – CONVERTED TEXTS

The following texts are those that have been extracted from CAIM's BergenBy database and used in this thesis. As many of the original stored texts were written in Norwegian, it was necessary to first translate them to English before they could be used in this thesis. The texts presented here will be displayed with its title/name as well as its corresponding ID used in the BergenBy database.

ID 1 – Beffen:

"Beffen" is a small ferry which takes passengers over Vågen in Bergen, more exact between Bryggen and Nordnes, about 50 times a day. Yearly there are between 25 and 30000 passengers that take Beffen. The boat has a 22.5 horse power engine and the voyage takes around 5-6 minutes. Many use the boat to get to work and school in the morning as some would take the bus. The name "Beffen" stands for Bergens Electric Ferry Company. In 1894 a local man by the name J. C. Troye took the initiative to start the company, and the first ferry was up and running August first the same year. In the beginning there were 4 small ferries put in 6 different liaisons in Vågen and Puddefjorden. Today there is only 1 ferry up and running between Bryggen and Nordnes. It is the twelth of its kind.

****

ID 2 – Bergen Natural History Museum:

The Bergen Museum is a university museum in Bergen, Norway. Founded in 1825 with the intent of building large collections in the fields of culture and natural history, it became the grounds for most of the academic activity in the city, a tradition which has prevailed since the museum became part of the University of Bergen. Bergen Museum is divided into two departments, the Natural History Collections and the Cultural History Collections. It is also the caretaker of the botanical garden surrounding the natural history building, and the cities arboretum. Bergen Museum was founded in 1825 by Wilhelm Frimann Koren Christie, at the time president of the Storting. In its early years, the museum contained numerous art collections, including several works by the painter Johan Christian Dahl, cultural artifacts, and craftwork items. In 1931, the museum moved from its location in the Seminarium Fredericianum building near Bergen katedralskole, to a new building south-west of Lille Lungegårdsvann. This was the first dedicated museum building in Norway. The current natural

history building was finished in 1865, and Bergen Museum moved in 1866. The botanical garden was laid out between 1897 and 1899, and the cultural history department got its own building in 1927. The increasing research activity at the museum from the late 19th century and onwards led directly to the founding of the University of Bergen in 1948.
****

ID 3 – Børs:
The old Børs building, today containing the Frescohall was completed in 1862, and expanded in 1893. The frescos are considered to be part of the Norwegian national heritage and were painted by Axel Revold (1887-1962). The frescos themselves were painted in the period between 1921 and 1923.
****

ID 4 – Christian Michelsen Statue:
The statue of Christian Michelsen (1857-1925) - created by Gustav Vigeland. At the Festplaussen's north-east side can be found the city's highest monument. Sucklelen is made of west-norwegian grannit and is 17.93 meters high, has his left hand in his pocket, and the other in a fist. After Christian Michelsen’s death the known sculptor Gustav Vigeland was asked to sculpt a monument in memory of the split of the union and Christian Michelsen's importance in connection to this. This caused strong reactions as it was standard procedure when raising a monument of national importance to arrange an artists’ competition. The monument was unveiled by King Håkon the twelfth on may seventeenth 1938, with the presence of both the "storting" and the government.
****

ID 5 – Edvard Grieg Statue:
The Edvard Grieg statue, unveiled the 4th of September 1917, is situated in the eastern part of Byparken, facing Musikkpaviljongen and Permanentent. Grieg is depicted leaning on his left foot, supported by a walking stick in his right hand. The left hand is holding on to his jacket pocket. His head is facing slighty upwards, and his hair is arranged delicately.
****

ID 6 – Festplassen:

Festplassen is the area between Lille Lungegaardsvann, Rasmus Meyers Allé, Christies gate and Kaigaten. It is used for the May 17th celebration (since 1929), fun fairs, amusement parks, feast day and festivals.

****

ID 7 – Fish Market:

The "Fisketorget", or fish market, in Bergen has over hundreds of years secured a position as one of Norway's most known and beloved outdoor- markets. The market has a charming site between western fjords and the seven mountains of Bergen, and with the combination of the smell of the sea one really feel like they are in Bergen. At the market it is first and foremost possible to buy or look at the most edible of things that come from the sea, but the fish market in Bergen also has many other products to offer. At the fish market, centered by "Bryggen" and the Zacharias brewery, the can be at times up to forty seller of wares of the highest variety. Tourists visiting the Hordaland region will most likely find live crab and lobster to be quite exotic, but at the market one can also buy many other wares and accessories. In addition to dishes with scampi, shell food, fresh fish, many types of smoked fish like salmon, mackerel and herring, not to mention the homemade fish cakes and Norwegian caviar, one can also buy authentic sausages of moose and dear. The fish market in Bergen has existed for a very long time. Already in 1276 was it decided by city laws where in the city it was permitted to bargain and trade wares. At this times the market was situated much further in the city, but as time went by more and more of "Vågen" was filled in. At the start of the nineteenth century, the fish market had been in danger of being completely shut down, but local powers and initiators managed to keep it standing. Today the fish market is a center of bustling trade and life in Bergen sentrum.

****

ID 8 – Gamlehaugen:

Gamlehaugen is the official royal estate in Bergen. The pompous building is situated in a parc ground between Fjøsanger and Hop, by Nordvannet in the Faana area. The castle like villa was registered for Christian Michelsen during the years 1900-1902 by drawings from architect Jens

Zetlitz Monrad Kielland, and was transfered to the state after Michelsens death. The parc is now open for visitors exept for when the royal family is in Bergen.

****

ID 9 – Grieghallen:

The Grieg Hall (Norwegian: Grieghallen) is a 1,500 seat concert hall in Bergen, Norway. It has been the home of the Bergen Philharmonic Orchestra since the halls completion in 1978. It hosted the Eurovision Song Contest in 1986, and is the host of the annual Norwegian Brass Band Championship competition, which occurs in mid-winter. The hall is named after Bergen-born composer Edvard Grieg, who was music director of the Bergen Philharmonic Orchestra from 1880 until 1882. The Greig Hall recording studio is also famous within the black metal community as several of the most important Norwegian black metal albums were recorded in this studio.

****

ID 10 – Johannes Church:

The Johannes church is a cross-church from 1894 in the Bergen community, Hordaland region. The church stands on Neegorsshayden in Bergen. It was built in the peroide of 1888-1894 drawn by the architect Herman Major Backer, and is with it's 61 meter above the ground Bergen's highest tower. It is also the church in Bergen with the largest amount of sitting space with room for 1 thousand two hundred and 50 people. It is built with brick in a new-gothic style and in the cross arm sections of the church there can be found "sideskip" and side-galleries. Fresco-paintings in the church's parish hall from 1924 are painted by Hugo Lous Mohr.

****

ID 11 – Lille Lungegårdsvatn:

"Lille Lungegårdsvannet" (Little Lungegårds Water) is an eight sided bed of water in the heart of Bergen city. Lille Lungegårdsvann is 700 meters in perimeter, and is not an artificial mere. Lille Lungegårdsvann was originaly connected to "Store Lungegårdsvannet" (Big Lungegårds Water), but were divided in 1926. On June twenty third, 2004, the city council leader, Monica Mæland, opened the new fountain. The fountain is now bigger than the old one and consists

of 21 nozzles and 36 spotlights which are meant to light up the "dance" of water at night. The old fountain paid it's toll as time went by and seased to function in 2003.
****

ID 12 – Long Curch:
The church is a long-church made of stone with side-wings and a "skrudhus". It is 60.5 meters lang and 20.5 meters wide. The towers width is 13 meters, while the quire is 13.5 meters wide. There exist ruins of the two older churches. The first time someone heard of them was in 1181, when it was called Olavskirken (Olavs church) in Vågsbunnen. It was also dedicated to Olav den Hellige (the Holy). Under Håkon Håkonsson's minitry/reign there was built a fransikaner closter a little south from the church which then became a part of it. The church burned down in 1248 and 1270. It could be used again in 1301 after Magnus Lagabøte had helped finans the restoration of it. When Magnus Lagabøte passed away it was here he was put to rest. Again the church burned down in 14 63/64, and was ruins when it in 1537 was rebuilt as "domkirke". The previous domkirke was Kristkirke on Holmen in Bergen which was torn down in 1531. Bishop Gjeble Pederson made sure of the re-raising and repairing of damages on the structure before his death in 1557. This tower was over the mid-wing and was replaced by a tower in the west-end around the 1640's when the church was given its present form. There were also exstensive restaurations after the great city fire in 1702, and the church was not used again until 1743. A larger rebuilding was conducted in 1880-1883 by architects Christie and Blix. The side wing was torn down and built from scratch, with 4 gothic columns. It was desired that the church be given back its middle-aged appearence and removed of its rococco interior. The church was then given a new alter in "kleber stone". The church also contains a memorial from the battle on Bergen's Våg between English and Hollanees ships in 1665. There still remains a canonball in the wall.
****

ID 13 – Ludvig Holberg Statue:
Located in the very heart of Bergen, by the tourist information office, you will find the monumental three metre high statue of Ludvig Holberg. This has long been one of the city’s key landmarks. It stands three metres high and is considered one of the finer works of Swedish sculptor Johan Boerjeson. The statue was unveiled in 1884. Ludvig Holberg lived

between 1684 and 1754. He was born in Bergen, although he lived most of his life in Denmark and died in Copenhagen. He is considered the founder of modern Danish literature.
****

ID 14 – Rogalands Cross:
The stone crucifix outside Fantoft stave church. Originaly one of four from Tjora outside Sola.
****

ID 15 – Rosenkrantz Tower:
The Rosenkrantz tower is in Bergen. The tower is named after Erik Ottesen Rosenkrantz who was the "lens-herre" of Bergenhus from 1559-1568. He had made extensive work on the tower. The tower's history is however much older than this. Two older constructions are built inside the tower. The oldest is king Magnus Lagabøte's castell which is a smaller tower from about 1520. This has led to the tower having dislocations in the floor hight. With its strategic placement in the south wing, the tower has been a cornerstone in the castle system in the Bergenhus fortress. The exsplosion in Vågen in 1944 led to extensive damages to the Rozenkranz tower, but it was later re-built.
****

ID 16 – Sailors Monument:
Ever since the unvealing in 1950, the twelve bronze sculptures and the four reliefs that make up the Sailors Monument have stood as symbols to Norway'sthousands of years history as a sea baring nation.
****

ID 17: Statsraad Lehmkuhl:
Statsroad Lehmkuhl is Bergens ``training ship´´. It was built as a rind in Bremerhaven in 1914 under the name Grossherzog Friedrich August as training ship for the german trade fleet. During World War One the ship was taken as a prize by Great Britain and the ship was in 1921 bought by former minister Kristofer Lehmkuhl. Since then, with the exception of World War Two the ship has been Bergens training ship.
****

ID 18 – Stave Church:

Fantoft stave church is a stave church in the area of Fana, in Bergen. It is a reconstruction of the church that burned down in 1992. The church was originaly built around the year 1150 and stood in Fortun, in Sogn. The work was lead by curator Anders Lorange. He surveyed the work when the church was taken apart and shipped by boat to Bergen. Carpenters from Sogn were hired and besides from Lorange, Joakin Mathiesen was an architect. Fortun-kirken had been changed a bit over time and had been given a new choir area from Lafteverk, and also had a west-tower from the 16 hundreds. However, in Fantoft all these addition had been neglected and the church was reconstructed the way people thought it had been originaly. The parts that were missing were also made the way people thought they once were with other stave churches, especially using the Borgund stave-church as models. The sixth of June 1992, the church burned almost to the ground. All that remained were portions of the "skeleton", and with no chance of rebuilding from the remains. Suspected in the case was Varg Vikernes. It was shortly after announced that a new stave church was to be built. Later a storehouse burned down containing remains that had been salvaged from the old church. The building of the new church was therefore a challenge for the mayor and carpenters as no stave church had been built for hundreds of years and the existing building knowledge was quite modest. Most of the building parts were also made at the site following measurements and existing drawings. Building of the church was put into motion and was finished in 1997. This was an identical copy of the church that burned down. Wood from Kaupanger had been used in the new church. The crusifix in the choir area is carved out by Sven Valvatn and painted by Solrun Nes. There was quite little that could heva been used from the old church, but a "wish stone", possibly a relic, which can be found in one of the walls, as well as the cross on the alter are said to be from the original church. It was also impossible to reconstruct wall paintings that were in the church before the fire. Outside of the church stands a stone cross from Tjora in Sola. The cross standing outside of the Bergen Museum is also from there.

****

ID 19 – Sæverud Monument:

The sculpture is a memorial sculpture of the composer Harald Sæverude (1897-1992). The piece of art consists of three round-bended sculptures in a rustic colored steel, and is a tribute

to Sæverude. The artist's idea has been that the circles should make us think of music. Originally, each circle had a unique surface; one in black satin, one in the color of rust and one in shining gold. The gold color was supposed to symnbolize the musical freedom of the composer. Because of constant cor'rosion under the paint it was decided in 2004, in cooperation with the artists, to remove the the painting. Even with constant sand blasting, rust treatments and regular painting, corrosion would have been a continuous problem. Today, the whole sculpture stands with an oiled rust. The sculpture raised debate when it was created. Some meant that a sculpture in memory of Harald Sæverud should be more naturalistic so that it stood equal to the other sculptures in the nearby area, while others meant that an abstract memorial was just what was needed and would "freshen up" and "renew" the surroundings. The french artist Bernar Venet has for many years workedconceptually with a classic modernistic approach. His massive sculptures, focusing on arcs and lines in the physical space, stand in central "free areas", or open spaces, around the world. In 1999 he won the competition for the Sæverud-monument in Bergen.
****

ID 20 – Troldhaugen:
Edvard Grieg Museum Troldhaugen was founded as Troldhaugen - Edvard and Nina Griegs home in 1928. The museum can be found in Hop in the Faana area in Bergen and consists of Griegs villa Troldhaugen from 1885, the Komponisthytten, Edvard and Nina Griegs grave, Kammermusikksalen Troldsalen from 1985, and a museum building from 1995. The museum has also administrated the Siljustøl Museum since 1997. From 2007 the Troldhaugen was consolidated with the the Lysøen Museum, the Western Art Industry Museum and the Bergen Art Museum under the name ``The Art Museums in Bergen´´.

APPENDIX E - RAW DATA FIGURES

In the tables presented in this appendix, columns marked in blue represent those of the participants that listened to clips in the order Small → Medium → Large. The values in the tables represent the answers given from participants, ranging from 1, being the most negative answer, to maximum 7, being the most positive answer. These "most positive" values vary between questions, and the maximum value for each question is shown in each table below as "levels 1-x" to the right of each quality.

Participants:		1			2			4			5			6			8			13			15			16			Avg
Gender:			F			F			F			F			F			F			F			F			F		
Duration:		L	M	S	L	M	S	L	M	S	L	M	S	L	M	S	L	M	S	L	M	S	L	M	S	L	M	S	
Audio Quality:																													
choppiness	levels 1-7	1	1	1	1	3	3	2	2	6	2	5	6	5	5	3	4	3	5	5	4	2	1	3	4	4	2	3	3,19
understandable	levels 1-5	3	4	3	2	2	3	4	2	4	4	4	5	4	4	4	4	4	4	4	3	4	2	4	4	4	3	2	3,48
tiring	levels 1-5	2	5	4	2	4	3	4	4	4	2	2	4	4	4	4	4	2	4	2	2	4	3	4	3	4	3	4	3,37
willingness to listen	leves 1-4	2	3	3	2	2	2	2	2	2	2	2	2	4	4	3	3	3	3	2	3	4	2	3	4	2	3	4	2,7

Figure E.1 – Raw data from female participants during computer-based testing

Participants:		3			7			9			10			11			12			14			17			Avg
Gender:			M			M			M			M			M			M			M			M		
Duration:		L	M	S	L	M	S	L	M	S	L	M	S	L	M	S	L	M	S	L	M	S	L	M	S	
Audio Quality:																										
choppiness	levels 1-7	1	1	2	5	2	3	3	3	5	3	3	2	3	3	3	3	1	6	1	3	3	1	1	2	2,63
understandable	levels 1-5	2	2	2	4	3	4	3	3	4	3	3	2	2	2	2	4	3	4	2	3	3	2	2	2	2,75
tiring	levels 1-5	2	2	4	4	2	4	2	2	3	2	4	4	1	1	1	4	3	5	1	4	4	2	2	3	2,75
willingness to listen	leves 1-4	1	1	1	3	2	2	1	1	1	2	2	2	1	1	1	2	3	4	1	3	4	1	1	2	1,79

Figure E.2 – Raw data from male participants during computer-based testing

Summary Data - Combined - Computer-Based Test					
		Avg	Min	Max	Mode
Audio Quality:					
choppiness	levels 1-7	2,92	1	6	3
understandable	levels 1-5	3,14	2	5	4
tiring	levels 1-5	3,08	1	5	4
willingness to listen	leves 1-4	2,27	1	4	2

Figure E.3 – Summary data from combined results of computer-based testing

Figure E.3 was generated by placing the average values from both male and female participants from the computer-based testing next to one another and calculating the presented sums.

Participants:		2_3			2_4			2_7			2_8			Avg
Gender:			F			F			F			F		
Duration:		L	M	S	L	M	S	L	M	S	L	M	S	
Audio Quality:														
choppiness	levels 1-7	1	4	4	5	5	3	4	5	6	3	5	4	4,08
understandable	levels 1-5	2	3	4	4	4	3	4	4	5	3	4	2	3,50
tiring	levels 1-5	3	4	5	4	4	4	2	4	5	3	3	4	3,75
willingness to listen	leves 1-4	3	4	4	2	3	3	3	4	4	2	4	4	3,33

Figure E.4 – Raw data from female participants during mobile-based testing

Participants:		2_1			2_2			2_5			2_6			Avg
Gender:			M			M			M			M		
Duration:		L	M	S	L	M	S	L	M	S	L	M	S	
Audio Quality:														
choppiness	levels 1-7	1	2	3	3	5	1	3	4	5	3	4	2	3
understandable	levels 1-5	2	3	3	3	4	3	3	4	4	3	3	3	3,17
tiring	levels 1-5	2	3	3	4	4	3	2	4	4	4	4	3	3,33
willingness to listen	leves 1-4	3	4	4	4	4	4	2	4	4	3	3	3	3,5

Figure E.5 – Raw data from male participants during mobile-based testing

Summary Data Combined Mobile-Based Test					
		Avg	Min	Max	Mode
Audio Quality:					
choppiness	levels 1-7	3,54	1	6	3
understandable	levels 1-5	3,32	2	5	3
tiring	levels 1-5	3,54	2	5	4
willingness to listen	leves 1-4	3,34	2	4	4

Figure E.6 – Summary data from combined results of mobile-based testing

Figure E.6 was generated by placing the average values from both male and female participants from the mobile-based testing next to one another and calculating the presented sums.

APPENDIX F – GRAPHS

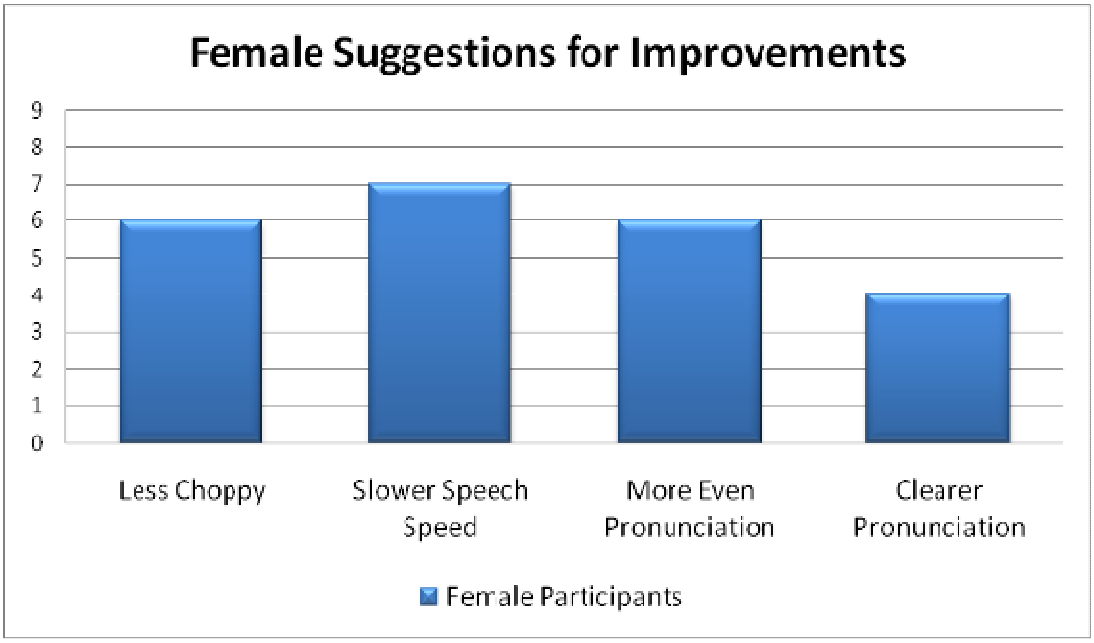

Figure F.1 – Diagram of female suggestions for improvements (computer-based testing)

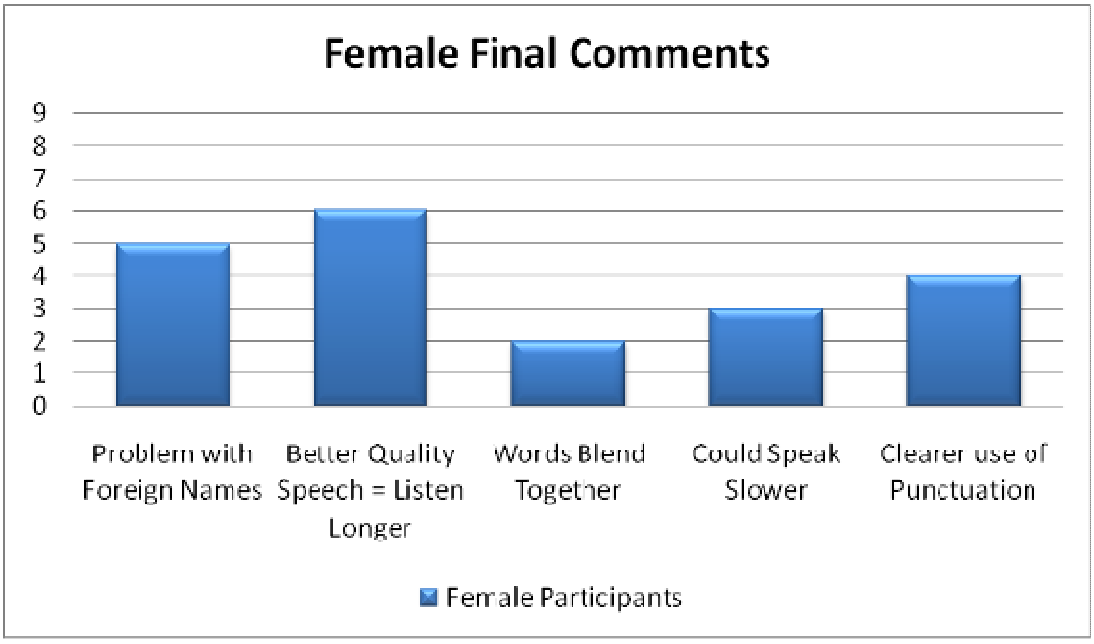

Figure F.2 – Diagram of final comments from female participants (computer-based testing)

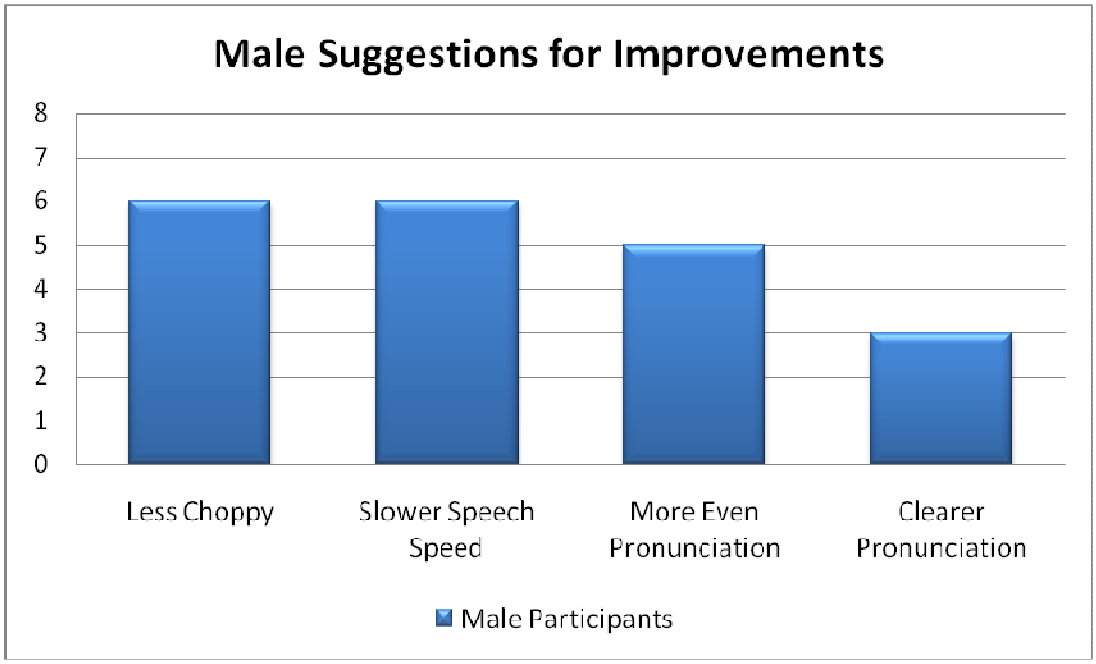

Figure F.3 – Diagram of male suggestions for improvements (computer-based testing)

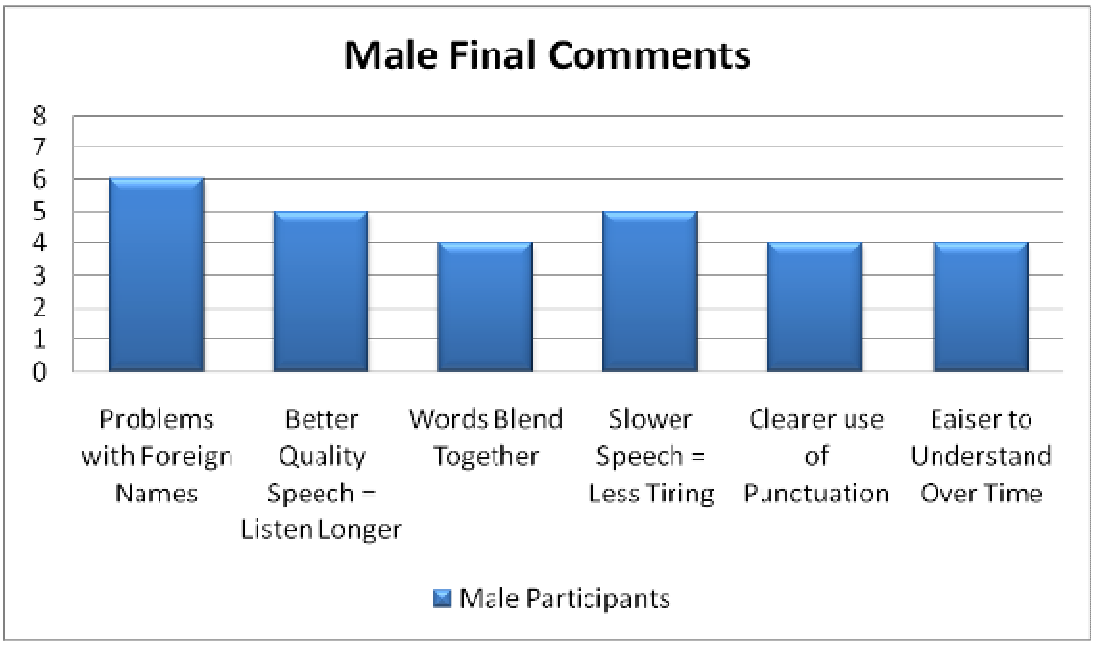

Figure F.4 – Diagram of final comments from male participants (computer-based testing)

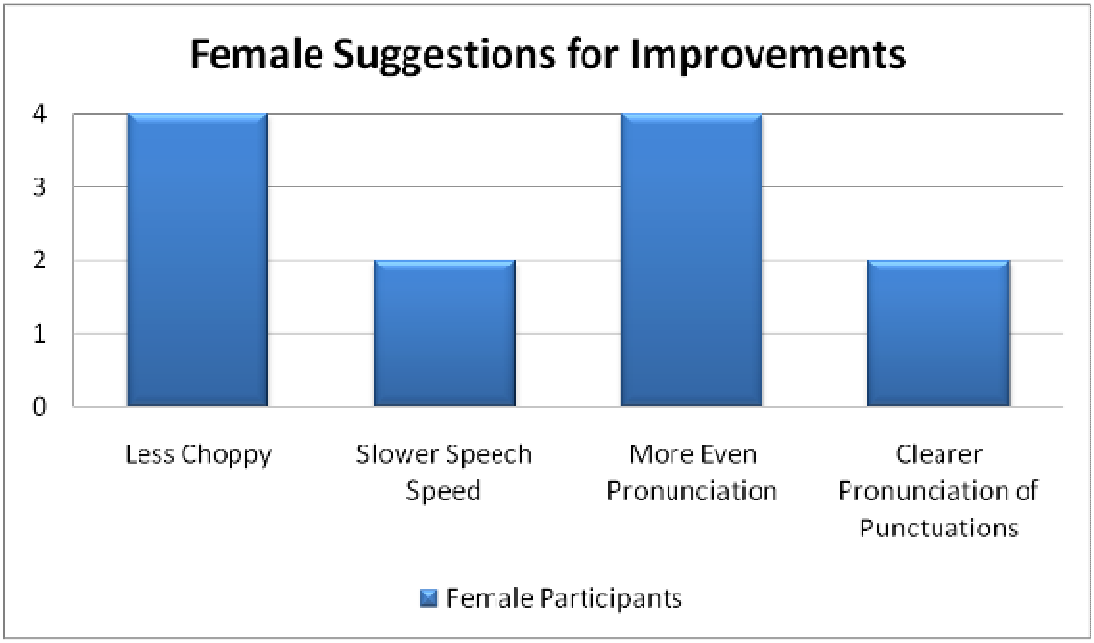

Figure F.5 – Female suggestions for improvements (mobile-based testing)

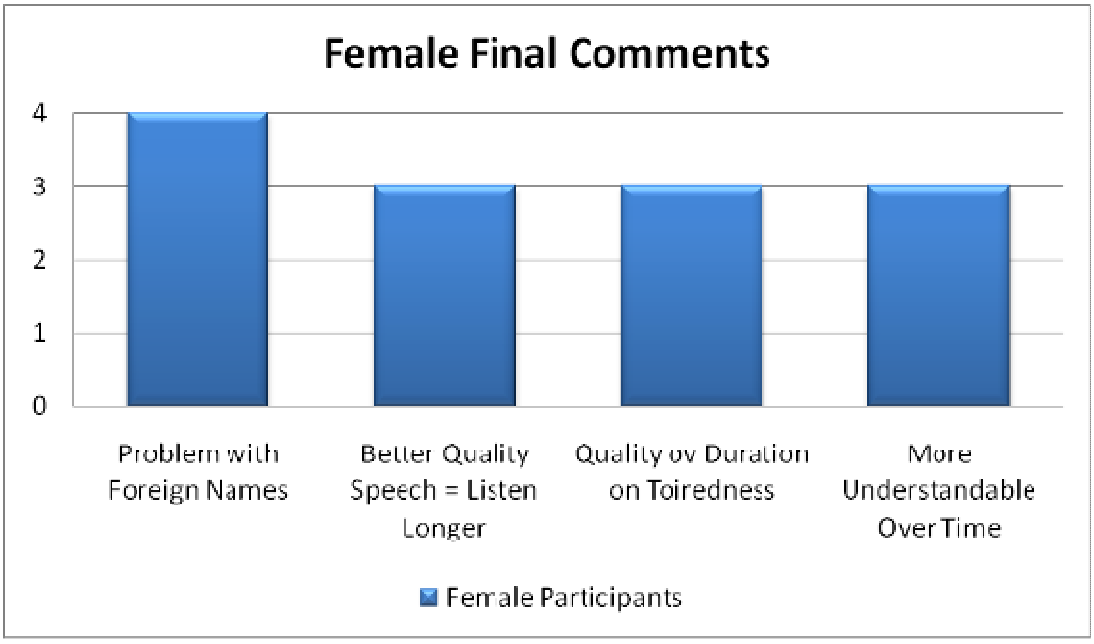

Figure F.6 – Diagram of final comments from female participants (mobile-based testing)

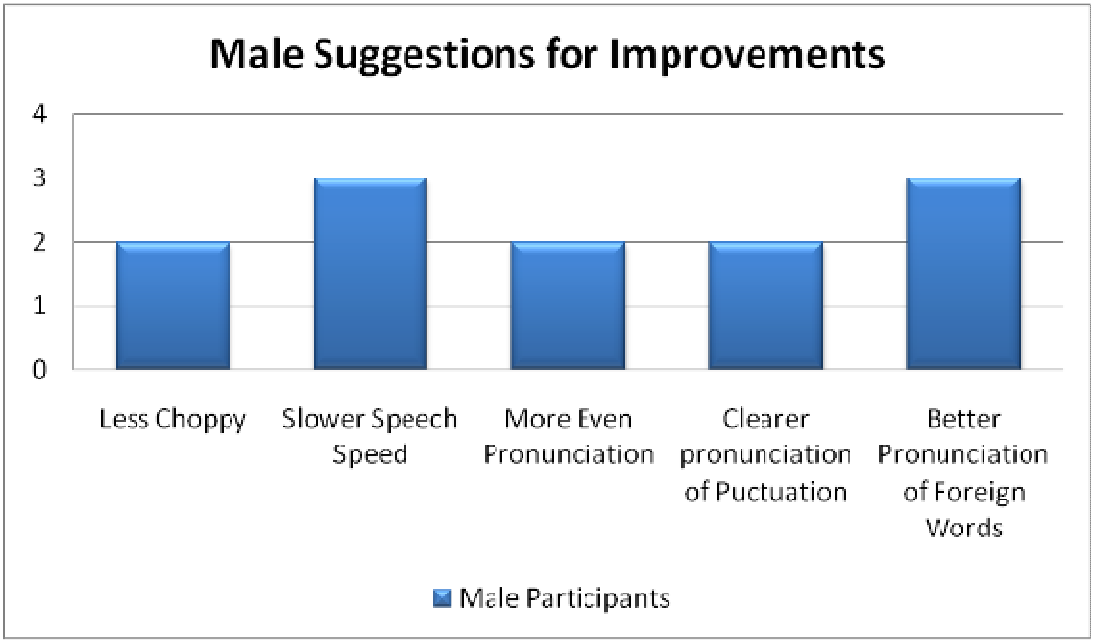

Figure F.7 – Male suggestions for improvements (mobile-based testing)

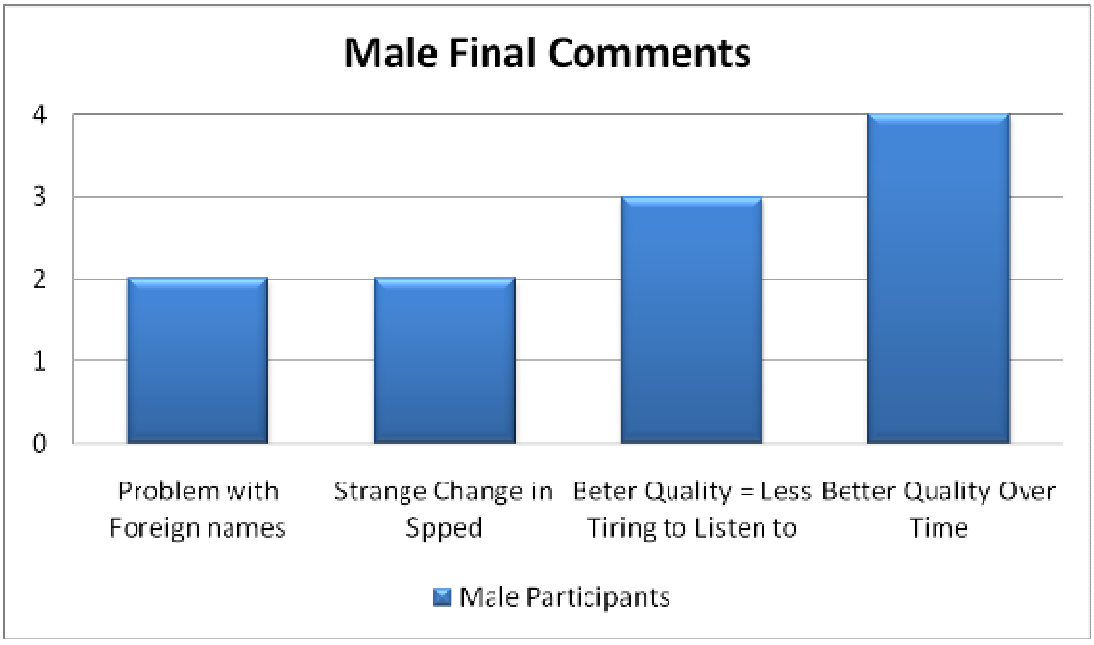

Figure F.8 – Diagram of final comments from male participants (mobile-based testing)

APPENDIX G – T TEST RESULTS

The following screenshot is of the results page given from the free online t test calculator used (GraphPad Softwares' homepage, 2009) using the following data sets:

(Group One)	(Group Two)
Comp-based summary	Mobil-based summary
2,92	3,54
3,14	3,33
3,08	3,54
2,27	3,34

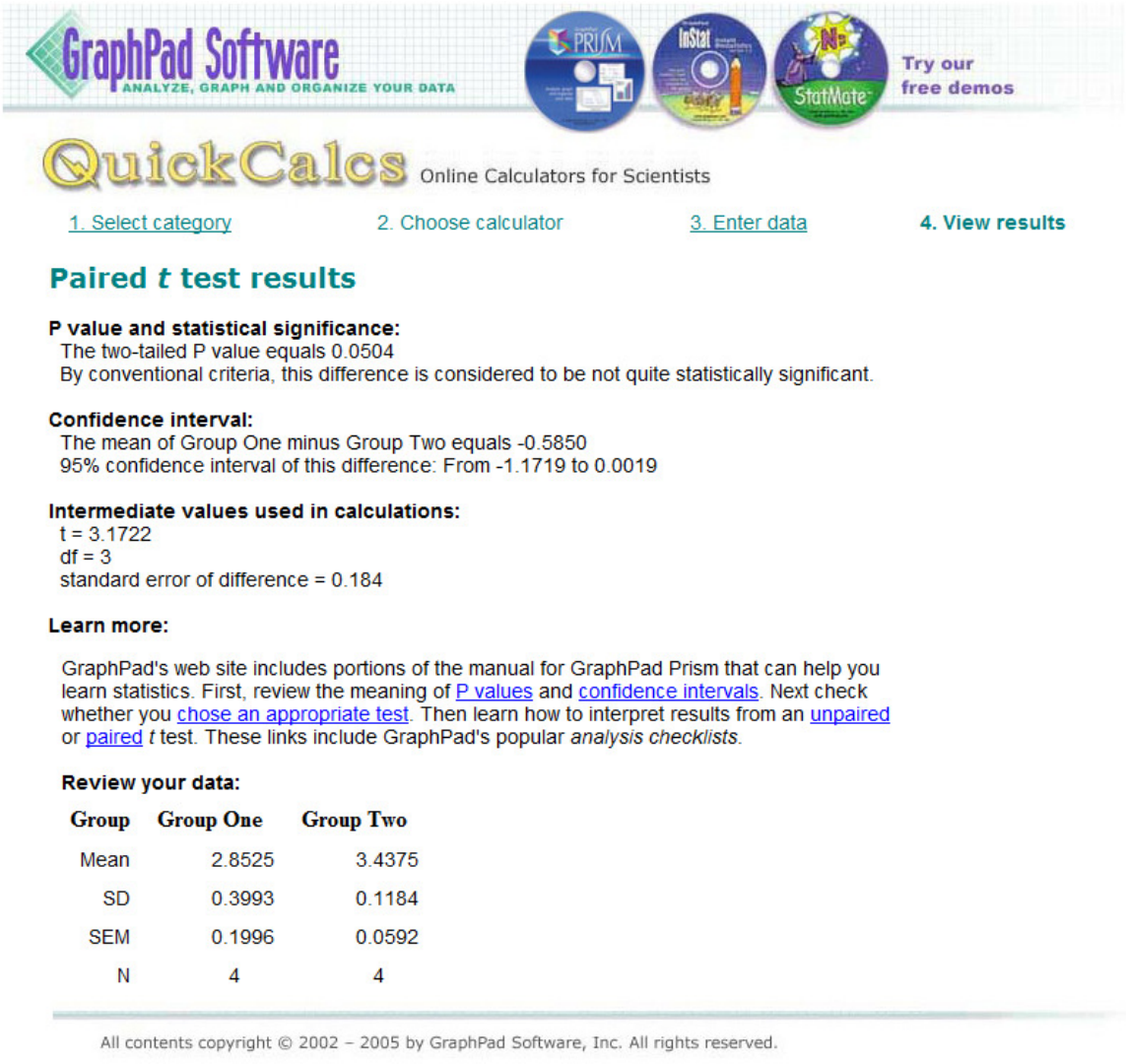

Figure G.1 – Screen shot of t test used for calculating statistical significance