Recording Methods for Disfluent Speech for Implementation of Conversational Synthesis Voices

MIRAY MERKES

Master of Science Thesis
Stockholm, Sweden 2009
Recording Methods for Disfluent Speech for Implementation of Conversational Synthesis Voices

M I R A Y M E R K E S

Master’s Thesis in Speech Communication (30 ECTS credits)
at the School of Media Technology
Royal Institute of Technology year 2009
Supervisor at CSC was Joakim Gustafson
Examiner was Rolf Carlson

TRITA-CSC-E 2009:099
ISRN-KTH/CSC/E--09/099--SE
ISSN-1653-5715

Royal Institute of Technology
School of Computer Science and Communication
KTH CSC
SE-100 44 Stockholm, Sweden
URL: www.csc.kth.se
“One of the most critical challenges is simply determining the speaker’s intended utterance”

By: Peter A. Heeman & James F. Allen,
(1999, p.527)
Abstract

Spontaneous speech is the spoken language that humans use when communicating with each other on everyday basis. It consists of both verbal and non-verbal speech. Verbal speech, in general, can be defined grammatically, syntactically and morphologically and causes therefore no bigger issues while implemented in spoken dialogue systems. The non-verbal speech, on the other hand, follows a less regular structure and their meanings changes depending on for example the intonation, voice-quality and loudness. Researchers around the world have done a great amount of work in the field of spontaneous speech and disfluencies. There is also a limited amount of studies that deal with the synthesis of disfluent speech. However, in this Master thesis, the science will be taken a step further. The goal is to find recording methods that give rise to spontaneous speech and that when synthesized the speech will sound natural. A second goal is to find a voice that is accepted by the naïve listener and that also fulfill the goals of sounding natural and being intelligible. This problem was approached by using two different methods; one was based on a human-human conversation that adapts the main idea behind the game DEAL (Klarenfjord, 2008) and the Swedish TV-show “Parlamentet” (Jarowskij et al., 1999). The second method was a laboratory read. Three recordings were done in total. Two of these were recorded in accordance with the first method using the theme on-line selling products and the third was based on read sentences that were taken from the two first dialogue recordings. About 20 sentences with disfluencies were chosen for the third recording. The interlocutor that participated in the third recording read first the 20 sentences in their original state, secondly where the disfluencies were put into new sentences and third were the new sentences were read without the disfluencies. The recordings were followed by transcription and synthetic generation of the recorded speech using Higgins Annotation tool for word transcription, Expres toolkit for the phone transcription and dipphone synthesis for the generation of the voices. Two main voices were used in the experiment; Ingmar which is a laboratory generated dipphone voice and Own which is based on the voice quality and the prosody of the interlocutor’s voice. In order to blur the direct distinctions between the original voice and the voice of Own, the original voice was encoded with AMR which is a wideband coding for cell phones. When all the recordings, transcriptions and synthetic generations were done the remaining was to test the collected data. This was done with a listening test that included 16 subjects and consisted of 5 parts. The first part included personal questions about the subjects whereas the remaining four parts tested and compared the different methods and different voices. When the results from the listening test were compiled and analyzed they showed that the voice encoded with AMR coding sounded most natural and Ingmar sounded the least natural, however all voices were good at communicating that a disfluency occurred. The results also revealed that disfluencies that are recorded in a dialogue give more spontaneous and natural sounding synthesis than does the played disfluencies (read sentences).
Sammanfattning

# Table of Content

1 Introduction ................................................................................................................................. 9

1.1 Purpose & method...................................................................................................................... 9

1.3 Overview of the paper................................................................................................................. 9

2 Theoretical Background ............................................................................................................. 11

2.1 Spontaneous speech.................................................................................................................. 11

2.1.1 Disfluent speech..................................................................................................................... 11

2.1.2 Why do interruptions in speech occur?.................................................................................... 13

2.1.3 Who contributes with the most disfluency?.............................................................................. 13

2.1.4 Advantages and disadvantages of disfluent speech................................................................. 14

2.1.5 Definition of terminology...................................................................................................... 14

2.1.6 Pauses..................................................................................................................................... 15

2.1.7 Truncations.............................................................................................................................. 16

2.1.8 Hesitations.............................................................................................................................. 16

2.1.9 Corrections.............................................................................................................................. 16

2.1.10 Fresh starts/False starts........................................................................................................ 16

2.1.11 Repetitions............................................................................................................................ 16

2.1.12 Non-verbal speech............................................................................................................... 17

2.2 Dialogue systems ..................................................................................................................... 18

2.2.1 How do a simple SDS work.................................................................................................... 18

2.2.1.1 Speech recognizer............................................................................................................... 19

2.2.1.2 Language understanding.................................................................................................... 20

2.2.1.3 Dialogue management....................................................................................................... 20

2.2.1.4 External knowledge base.................................................................................................... 21

2.2.1.5 Language generator/ response generator.......................................................................... 21

2.2.1.6 Text-to-speech synthesis.................................................................................................... 21

2.3 Synthesizer technologies ......................................................................................................... 22

2.3.1 Concatenate synthesis............................................................................................................ 22

2.3.2 Formant synthesis.................................................................................................................. 23

2.3.3 Hidden Markov Model (HMM)............................................................................................... 24

2.3.4 Synthesis of disfluent speech................................................................................................. 24

2.4 Transcription alphabets ............................................................................................................ 26

2.4.1 IPA......................................................................................................................................... 26

2.4.2 SAMPA................................................................................................................................... 26

2.4.3 STA......................................................................................................................................... 26

2.5 Methods for data collection ..................................................................................................... 27

2.5.1 Blaue, Eileen & Kay................................................................................................................ 28

2.6 DEAL ......................................................................................................................................... 29

3 Method ......................................................................................................................................... 30

3.1 Data collection.......................................................................................................................... 30

3.1.1 Scenario design........................................................................................................................ 30

3.1.2 Recordings............................................................................................................................... 34

3.1.2.1 Recording 1......................................................................................................................... 34

3.1.2.2 Recording 2......................................................................................................................... 36

3.1.2.3 Recording 3........................................................................................................................ 37
1 Introduction

Spoken dialogue systems (SDS’s) have existed for a few decades now, but the systems still have problems understanding as well as generating conversational speech. In this paper an attempt is done to solve the second part of the problem, meaning producing natural and intelligible synthetic voices. The problem will be approached in an early stage, namely at the recording step of the process. Spoken dialogue systems are created for the naive user and therefore when developing the recording methods and the synthetic voices the user is put in the centre.

1.1 Purpose & method

The main purpose of this master thesis is to investigate the types of recording methods that are most suitable for provoking disfluent speech. The goal is also to find a recording method that when the collected data is synthesized it will produce synthetic voices that sound natural and are intelligible.

Spontaneous speech is characterized by interruptions in the flow of speech which are called disfluencies. Hesitations, pauses, repetitions, repairs, and truncations, are some examples of different disfluency types in speech. How these interruptions in speech sound when synthesized and whether or not they are identifiable by the naïve user are some of the questions that this project will try to answer.

In this study three hypotheses were investigated:

1. Synthesis generated for use in spoken dialogue systems sounds better if they are based on speech corpora provoked in a dialogue situation rather than in laboratory read.
2. Synthesis of disfluent speech can be generated by using two methods; speech which is recorded in a dialogue or by played disfluencies (laboratory read).
3. Disfluencies can be elicited by putting up a role-play that is performed by two individuals, whom are restricted to a specific domain. The elicitation is performed by a third person that eavesdrops on the conversation that takes place between the two role-players and sends them instructions.

In order to collect the data needed to test the hypothesis of the current study two recording methods were put to test, one that adapted a role-play similar to DEAL (Klarenfjord, 2008) and a second method that was based on reading dialogue sentences from a manuscript. All together, data was collected from three different recording sessions, where two of those were performed in a dialogue. Three different participants were included in the recordings; two females and one male. The male candidate participated in all three sessions. A listening test was performed to investigate whether or not the disfluencies were identifiable or not when synthesized and also what type of method that gave rise to more natural synthesized disfluencies. Moreover, different diphone voices were compared against each other in order to find the most intelligible voice.

1.3 Overview of the paper

The paper starts by giving a theoretical background of the subject spontaneous speech and spoken dialogue systems (SDS) in chapter 2. This is followed by a description of the method
in chapter 3. In this chapter the methods for both the recordings and the listening test are presented to the reader. In chapter 4 the collected data is presented and analyzed and thereafter discussed in chapter 5. A conclusion is drawn in chapter 6 and some refinements for future research are given in chapter 7. The report ends with acknowledgments (chapter 8), followed by a reference list (chapter 9), a wordlist (chapter 10) and appendix (chapter 11).
2 Theoretical Background

Spoken dialogue systems are used worldwide and their technology has advanced excessively, however the implementation of spontaneous speech still remains unsolved. Current systems can neither speak spontaneously nor understand spontaneous speech. The main problem with spontaneous speech is the occurrence of interruptions and disfluencies that disrupt the flow of speech. Researchers, though many studies are done in the field of speech productions, do still not grasp how to define disfluencies in order for the system to distinguish them from fluent speech. In this section the work and conclusions of researchers’ will be put to light in order to give an idea of the problem that will be dealt with in this project.

2.1 Spontaneous speech

Spontaneous speech is the language humans use when communicating with each other on a daily basis. Natural speech differ from written text in the sense that it primarily lacks the presence of punctuation and secondary a big part is presented by disfluent speech (speech that does not follow the syntax rules of a language nor its grammatical foundation) (Kahn and Ostendorf, 2005). Since disfluent speech represent a big part of the verbal communication it is proper to be able to synthesize it. Further, it is considered to contribute to the naturalness of the voice and understandability of the unstructured text (Adell, Bonafonte and Escudero). Robert Eklund (2004, p.38-39) defines continuous speech as unrehearsed, real-time and on-line speech. These conditions under where the spoken word takes place contribute to the well-known phenomena called disfluency in speech (also referred to as interrupted speech). There are no precise numbers of the rate of disfluency in natural speech, but a rough estimation has been done to 5-10 % (Shriberg, 2001, p. 155; Eklund, 2004, p. 39).

2.1.1 Disfluent speech

“Different types of disfluencies arise from different underlying factors”

(Shriberg, 2001, p. 156)

Disfluences interrupt the flow of speech and hence, they add nothing to the propositional content of the utterance (Adell et al.; Bock, 1996, p.402; Corley and Hartsuiker). However while not contributing anything to the propositional content they are very important for communicating attitude and when wanting to keep the floor or emphasize something. Disfluent speech consists of hesitations, fillers, pauses, repetitions, repairs, truncated words (cutoffs in mid-utterance) and false starts. Truncated words can either be caused by the speaker him-self or by other surrounding factors such as a listener. According to Maclay Osgood’s general taxonomy (1959) repairs and truncated words aren’t classified as disfluent speech. If studying the work of Shriberg, on the other hand, it’s found that the author mainly mentions filled pauses (in other texts they can be referred to as hesitations), repetitions, repairs and false starts as examples of reoccurring disfluencies in speech. However, in this report disfluent speech will include hesitations (filled pauses), silent pauses, truncated words, repetitions and repairs. Disfluent speech, for most part, lacks any specific grammatical construction and occurs within the flow of fluent speech. Their prosodic and phonetic set up differs from that of verbal speech and they are related to the environment to where the speech takes place (Shriberg, 2001, p. 160-167; Shriberg, section 1; Nigel Ward, p. 1). While reading Eklund’s essay (2004, p. 43, 51-55) the reader discover that disrupted speech may have many different perspectives. In complement to the disfluency types mentioned earlier the writer add
stuttering and Meta speech to the list of interruptions in speech. Meta speech is sounds such as laughter, coughing, inhalation, exhalation and clearing of voice. Further, Eklund (2004, p. 57) refer to the work of Johnson et al. in both 1948 (p. 180-181) and in 1959 (p. 134-135). In the former work, Johnson et al. define disfluency as following:

- Repeated sounds, syllables, words or phrases
- Prolonged sounds
- Pauses
- Blockages (truncated words)
- Hesitations and
- False starts

Here the author divides repetition in speech into different subtypes, but in this present study they will be included under the same title. In addition, Johnson et al. define prolonged sounds as a separate type whereas in this work they will be referred to as hesitations. In the example of 1959, also known as “Johnson’s eight categories”, the following classification was done;

1. Interjections of sounds, syllables, words and phrases
2. Part-word repetitions
3. Word-repetitions
4. Phrase-repetitions
5. Revisions (= repairs)
6. Incomplete phrases
7. Broken words (truncated words)
8. Prolonged words

Interjections express emotions and are for the most part not connected to the grammatical foundation of the surrounding speech (Mcfadyen, H.). Fillers and words or expressions followed by an exclamation mark are generally called interjections. Put also notice to how the authors have sliced repetition in three separate categories likewise truncation into two categories (incomplete phrases and broken words). In other words, this classification can be shortened to five categories;

1. Interjections
2. Repetitions
3. Revisions
4. Truncations
5. Prolonged words

The number of different classifications is endless, but at the end they describe the same phenomena. Shriberg (2001, p. 155-table1), for example, divides disfluencies in six different types;

1. Filled pauses
2. Repetitions
3. Deletion
4. Substitution
5. Insertion
6. Articulation error

Deletion and substitution, insertion or articulation error can in other works be referred to as false starts and repairs respectively.
2.1.2 Why do interruptions in speech occur?

Interruptions are part of the daily spoken language. The reason for why humans take aid of disfluencies when speaking or trying to explain or express something have, in fact, many diverse explanations. In the study that M.F Garret did in 1982 (Kathryn Bock, 1996, p. 402) lead the researcher to the conclusion that discontinuities occurred because of for example the inaccessibility of finding the right expression at a specific moment, planning and retrieval of speech units and/ or speech processing. Another research that supports these assumptions is the work of Carletta, Caley and Isard (p. 1-3), which deals with the human production of language under time constraints. Even in this case the gathered corpus revealed that disfluencies occur due to the fact that the speaker;

1. Does not find the right expression
2. To win time before deciding what to say
3. Signaling that they are preparing to speak or
4. Stalling to keep the floor

2.1.3 Who contributes with the most disfluency?

Shriberg (2001, p. 154-159) implies that the rate of disfluency is very much dependent on the environment where the spontaneous speech takes place and also on the gender of the speaker and the gender of the conversational partner (the listener). In the paper the author reported that the observations that were done during the experiment lead to the conclusions that male speakers contribute with a higher rate of disfluencies. In addition, a male speaking to another male or a female speaking to a male will result in a higher rate of discontinued speech than would a conversation between two females. Further, Shriberg (2001) implies that the position of the disfluency and the length and complexity of the sentence also play an important role when deciding the rate of disfluency in a conversation. Other observations done in this experiment were that the human-computer conversation contained less disrupted speech compared to the human-human conversation and whether the dialogue was goal-oriented or not had no effect on the rate. Humans have the tendency to adjust their speech according to the listener’s comprehension. Meaning, when a person speaks to a non-native partner that has poor knowledge about the spoken language they tend to speak slower and more clearly. They pronounce the words more carefully. Humans have the ability to adapt their speech to the skills of their conversational partner (Gustafson et al., 1997, p. 4; Klarenfjord, p.13). The same conditions apply for the human-computer conversation. Computers, till this day, speak and understand a language that follows the rules of grammar rather than the spontaneous speech that humans use while communicating with one another. In an attempt to make them-selves understood, humans exclude utterances and phrases that does not have a real grammatical or verbal structure i.e. disfluencies. And this can be the reason for why human-computer conversations contain lower rates of disfluencies than does human-human conversations.

According to Shriberg (2001, p. 157) the speakers can be divided into two groups;

1. Repeaters and
2. Deleters

Repeaters are often characterized by being slow speakers, which in turn produces many hesitations and repetitions. Deleters, on the other hand, tend to speak very fast and therefore they often restart an expression or a phrase. This phenomenon is usually referred to as false starts.
2.1.4 Advantages and disadvantages of disfluent speech

The advantages of disfluencies in speech are that they;

- Have short length but big impact on letting a conversation go on
- Emphasis a specific utterance, phrase or word
- Ease the conversation
- Don’t intervene with the content of the expression nor changes the meaning of the word or sentence
- Add more flavor to the conversation
- Give a confirmation or a disconfirmation without interrupting the ongoing conversation
  (Ward; Adell et al.; Bock, 1996)

The disadvantages of disfluencies in speech are that they

- Are not easy to detect
- Cause problems in speech recognition
- Can lead to wrongly processed speech when misrecognized
- Are difficult to apply in Spoken Dialogue Systems due to their non-restricted acoustic features
  (Campbell, 2007; Ward; McTear, 2004)

2.1.5 Definition of terminology

Before moving on to the next topic of disfluencies there is a need for defining a few terms. When humans speak they often tend to interrupt themselves in case of a misused expression or a misused word. When interrupted they correct or modify what they have just said and move on. The part where they interrupt themselves and add the modifications to the speech is called speech repairs and according to Hindle 1983 as referred to in (Heeman and Allen, 1999, p. 529-530) this phenomenon can be divided into three categories;

1. Fresh starts
2. Modifications
3. Abridged repairs

Fresh starts mean that the speaker abandons the current utterance and starts all over again. Fresh starts occur either in the editing term or at the onset of the alteration. Moreover, there is little or no correlation between the reparandum and the alteration. See example 1 for clarification;

Example 1;

I want to by **ehh** How many pants do I afford to buy?

In modification repairs the speaker modifies, or correct, what he/she just have said. Sometimes a person says something but had the intention to say something else. When he/she detect that he/she has used the wrong expression he/she tends to interrupt him-/her-self and exchange the wrong uttered word with the right one. In this case there is a strong word correspondence between the repair and the alteration since the former utterance is not abandoned totally, instead a specific part of the utterance is modified.

1 Interruption point is the point where an interruption in speech is done for modification.
Example 2;

Put the box of oranges of apples in the kitchen

Abridged repair is a repair that lacks the presence of both a reparandum and an alteration, meaning that the phrase only consists of an editing term.

Example 3;

Can we uh take the bus to school?

Speech repairs are divided into three intervals (Heeman et al., 1999, p. 528);

1. Reparandum is the part of the speech which is being replaced
2. Editing term can be utterances such as filled pauses and cue phrases (Ballantine, J., 2005)
3. Alteration is the part of speech that is replacing the reparandum

Example 4;

Can we by two boxes of bananas no I mean two boxes of oranges

2.1.6 Pauses

A speaker sometimes pauses between two phrases or two words before continuing and these pauses can be of two kinds, filled or silent. A silent pause is when the speaker pauses without uttering anything. Filled pauses, on the other hand, are pauses where the speaker fill out the silence with non-verbal speech such as um, ehh etc. Filled and silent pauses are often referred to as hesitations since the speaker hesitate until he/she finds the expression they were looking for. Utterances like um and ehh don’t add anything to the meaning of the sentence, they only work as stallers until the speaker is ready to speak again.

Pauses are good indicators of speech interruptions and they signal the start of a repair (Howell and Young, 1991, p. 736). Further, the experiments done by Howell et al. (1991, p. 757) showed that a pause before a retrace aids the listener in understanding the message in opposite to a repeat. Adell et al., with accordance of the results gained by Sundaram and Narayanan (2003), did a study about analysis of disfluent speech and synthesis. Their findings imply that fillers aid the listener in understanding the text at the same time as it eases the flow of the speech. Not enough with that, they are easy to generate since silent pauses can be introduced before and after the interruption which avoid co-articulation problems.

Pauses combined with stress are good features to highlight the interruptions of speech and at the same time what has been altered. According to a study done by Carletta et al. (p. 2), pauses are a sort of strategy that the speaker uses in order to keep the floor. Boomer in 1965 as referred to in (Bock, 1996, p. 402) discovered that speakers utilize pauses mainly for three reasons. The first reason is to gain time to think about what one wish to say before doing so. The second reason is to plan the next constituent and thirdly to delay the retrieval of the next word to be spoken. Moreover, Deese in 1980 as also referred to in (Bock, 1996, p. 402) reported that pauses are more likely to occur in planned speech than unplanned.
2.1.7 Truncations

Interruptions are actually stops in the flow of spontaneous speech. Pauses, hesitations, false starts, repetitions etc. are all different kinds of interruptions. These are all interruptions caused by the speaker him-self, but there is a second kind of interruption which is caused by the opposite part. When the listener sometimes tries to take the floor he/she often needs to interrupt the present speaker by starting to speak. In this case the flow of the first speaker is interrupted to give place to the new speaker. In this study, however, the main focus will be on interruptions caused by the speaker him-self. Truncations, in general, are incomplete phrases, words or expressions (Rabold, 2002).

2.1.8 Hesitations

Hesitations are, as mentioned in section 2.1.6, defined as a combination of both filled and silent pauses. In 1982 Garrett as referred to in (Bock, 1996, p. 402) observed that hesitations reflect temporary increase in processing load. They are used to gain time for planning and retrieving the next step or as a delay till the speaker finds the expression or information he/she was looking for.

2.1.9 Corrections

Corrections are as the name indicates corrections of what the speaker just have said. Sometimes when speaking fast humans tend to utter the wrong color or the wrong name etc. And in the case of the occurrence of this phenomenon the speaker needs to pause, change the wrong word with the right one before being able to proceed forward. According to Deese in 1980 as referred to in (Bock, 1996, p. 402) corrections are more expected to occur in unplanned speech than in planned speech.

2.1.10 Fresh starts/False starts

In section 2.1.5 repairs and different categories of repairs were discussed. If recalling, one of these repairs was classified as fresh starts. Fresh starts, as explained earlier, are defined as that the speaker abandons the current speech and starts from “zero”. In order to have a fresh start one need a false start, since false starts are the substituent utterances for what the speaker wish to say but have difficulty in finding the expression for that moment. Therefore speaker uses the next close expression that expresses their intention. False starts occur often in unplanned speech (Bock, 1996, p. 402).

2.1.11 Repetitions

Repetitions are most likely to occur in unplanned speech according to Deese (1980) as referred to in (Bock, 1996, p. 402)). Repetitions are used by speakers to signal to the listener that they are still preparing to speak. This is especially emphasized when the speaker uses flat intonation, since flat intonations do not challenge the repeated utterance (Carletta et al., p. 3). Other areas where repetitions are used are when the speaker needs more time to formulate an expression, to stress a specific part of the speech and/or to restart a sentence (Adell et al.). Adell et al. found that generated repetitions are often perceived as errors in the synthesis rather than disfluencies. Repetitions are also called redundant in speech since they do not contribute with anything to the essence of the sentence.
2.1.12 Non-verbal speech

Spontaneous speech, as indicated in the earlier sections, is characterized by disfluent speech, or interruptions in the flow of speech as many wish to call it. However, what is not really emphasized is that natural speech is divided into two categories;

1. Verbal speech and
2. Non-verbal speech

Verbal speech follows a grammatical, syntactical and morphological structure. The non-verbal speech, on the contrary, does not. Non-verbal utterances add nothing to the propositional content of speech, instead they work as indicators that signal different messages to the listener. They also characterize the generated voice with naturalness when concatenated (Sundaram et al., 2003; Campbell, 2007, p.22). In table 2.1 below the non-verbal speech is divided into three categories, where different sounds are classified differently (Sundaram et al., 2003; Campbell, 2007, p.24).

Table 2.1: Non-verbal words divided into 3 categories.

<table>
<thead>
<tr>
<th>Paralinguistic cues</th>
<th>Disfluency</th>
<th>Reflexes</th>
</tr>
</thead>
<tbody>
<tr>
<td>Falsetto</td>
<td>Fillers</td>
<td>Throat clearing</td>
</tr>
<tr>
<td>Creaky</td>
<td>Repetitions</td>
<td>Lip smacking</td>
</tr>
<tr>
<td>Laughter giggle</td>
<td>Hesitations</td>
<td>Breathing</td>
</tr>
<tr>
<td>Sob</td>
<td>Pauses</td>
<td>Sniff</td>
</tr>
<tr>
<td>Cry</td>
<td>Repairs</td>
<td>Tongue clucking</td>
</tr>
<tr>
<td></td>
<td>Cutoffs</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Words like okey, oh, so etc.</td>
<td></td>
</tr>
</tbody>
</table>

The problem with the non-verbal speech is that depending on the intonation, voice quality, loudness and timing the intention changes, which causes problems when generated since there is no specific parametric value that characterizes these phenomena. Therefore, more attention and time should be devoted to the study of the prosodic features and voice qualities of non-verbal speech for synthetic generation (Campbell, 2007, p.24).
2.2 Dialogue systems

Dialogue systems are systems that are able to engage in a dialogue with a human no matter if the communication tool is speech-, text- or picture based or a mixture of them. There are many different types of dialogue systems and these differ in complexity level. Two examples are the Swedbank’s telephone support (0771 22 11 22) and Telia’s tele support (90200). Both supply customer services over the phone but with the difference that Swedbank’s support service expect answers by choosing the alternative that best suites the customer’s question by pressing a button or several. In tele support, on the other hand, the caller speaks to the system and hence a bigger database is required since the answers are stored to avoid repetition. Swedbank’s support service is less sensitive for bias since the only speaker is the system itself. Telia’s support service, on the contrary, is very intolerant to hesitations and pauses. It doesn’t give the customer enough time to choose the most appropriate expression for the problem. The system informs the user when it fails to recognize an expression. In overall the system functions well as long as the user keeps the answers short and concise.

As mentioned earlier dialogue systems can be built with different complexity levels, all depending on the purpose and goal of the project. The simplest form is a system that asks questions and requires short answers in return. The answers given are not stored and therefore no memory capacity is needed. The more complex versions require a big database since the answers given by the user are stored to avoid repetition. In addition, the system has the ability to verify a state and continue giving help from there. For example, in a travel instruction system the user may have come half way to the final destination without any aid, but would from now and forward need some help to continue. Here the system needs, primarily, to understand the problem, then verify how far the driver has come and finely, when all that is set, give the right instructions to the user. (Jönsson, 2007)

2.2.1 How do a simple SDS work

A dialogue system can be constructed to be system-initiative, mixed-initiative or user-initiative. A system-initiative means that the system takes the command and in a user-initiative is when the user takes the command. In a mixed-initiative system both the user and the system act as commanders (McTear, 2004, p. 108; Klarenfjord, 2008, p. 16 ff).

Spoken dialogue systems are as the name indicates systems that use voice as the main communication tool. The complexity level of a dialogue system depends on the main task of the system. The simplest version is systems that give information, such as Google 411 or a bank. The most complex version is in the domain of entertainment and chatting. A spoken dialogue system consists of six main components;

1. Speech recognizer
2. Text-to-speech (TTS)
3. Response generator
4. Dialogue manager
5. Knowledge base
6. Natural language understanding
When a dialogue is initiated between a system and a speaker information will be exchanged back and forward between the two. In order for the system to understand and respond on what have been said to it, it must primarily recognize the speech spoken by the other speaker. This step is referred to as speech recognition. The second step of the process is language understanding and here the speech is given a meaning. In the step of dialogue management two tasks are performed. The first one is to determine how the speech fits into the dialogue and the second is to decide on where to go from here. In the forth step, which is called knowledge base, information is retrieved from different resources in order to give an answer (a response). In language generation adequate words and phrases are chosen to be used in the response. Finally, the response is spoken through text-to-speech synthesis.

**2.2.1.1 Speech recognizer**

A speech signal is continuous. When entered in the speech recognizer device it is converted into discrete signals (words, phrases etc.). These discrete components are later compared to the available vocabulary in the database of the system and the best matching value is returned. However, due to external and internal factors the recognition can very often fail. The problem with speech recognition, especially in spontaneous speech, is that the boundaries are not clear. The speech is composed by disfluencies and verbal and non-verbal speech, coarticulations, noisy environments etc. Additional factors that impose problems on speech recognition are the inter-speaker variability and the intra-speaker variability. The former factor say that depending on the speaker’s gender, vocal track, origin etc. the same word can be pronounced differently by different speakers. The second factor indicates that some deviations in pronunciation of the same word at different occasions is possible, since the voice of a specific speaker can vary depending on for example the mood of the speaker and the speaker’s health (different acoustic properties of the same word and by the same speaker at different occasions).
Parameters that are necessary to deal with in speech recognition and that can facilitate or complicate the recognition of speech are:

1. **Enrolment;** speaker-dependent (trained for a specific speaker) or speaker-independent (trained for a wide range of users that are representative for potential users - more complex).

2. **Vocabulary size;** larger vocabulary give more flexible but more complex systems. A small vocabulary, on the contrary, is very easy to use but is also very restricted in use. Are often domain-specific.

3. **Speaking mode;** systems based on discrete speech are very easy and uncomplicated as long as the user pause between each word. Systems that are able to handle continuous speech are more complicated since they have to segment the speech and that is a difficult process because of the vague boundaries. The advantage on the other hand is that the user can speak freely without any pauses and clear articulations.

(Hjalmarsson, 2002, p. 11-12; McTear, 2004, p. 81-88)

### 2.2.1.2 Language understanding

In language understanding the output from the speech recognizer is analyzed and assigned a meaning that can be further processed by the dialogue manager. Language understanding is divided into two steps;

1. **Syntactic analysis and**
2. **Semantic analysis**

In the syntactic analysis the system studies the structure of the grouping of the words, whereas in the semantic analysis it is determined why these words are grouped in this specific way. Problems that can arise during language understanding can depend on;

1. Ambiguity (verbal ambiguity, sense ambiguity, structural ambiguity)
2. Illformed input

(Hjalmarsson, 2002, p. 12; McTear, 2004, p. 91-92)

### 2.2.1.3 Dialogue management

The dialogue management is the most important step in a dialogue system since it controls the dialogue flow. The task of a dialogue manager is, in short, to accept input from a speaker, process the input and later produce a message to be delivered to the user. In case the dialogue system does not find what it is looking for it asks an external knowledge source (knowledge base). Thus, a dialogue manager performs the following actions;

- Information retrieval
- Contextual understanding
- Determine whether or not enough information is gathered
- Generate a response

In case of misunderstanding or insufficient information the system uses verification and clarification strategies, such as asking the end user to confirm a specific statement. The complexity levels of a dialogue manager can differ depending on several factors;
• The level of the spoken language: Is the vocabulary constrained or does the user have the freedom to speak freely?
• Dialogue initiative: Who is leading the conversation, the system, the user or both (see section 2.2.2)?
• Control over the dialogue flow: Scripted dialogue means that the questions and the order they are asked in is fixed. Scripted systems are usually used in system-led dialogues. The open-ended systems have a more free approach. Here the next step, or question, is determined by the current state and therefore give way for more flexible dialogues.
• Grounding: This part deals with verifying that what has been said have been mutually understood. Further, if the user gives unclear or ambiguous information the system asks for clarification.


2.2.1.4 External knowledge base

External knowledge sources can be databases that contain information about what the user is asking for. The dialogue manager retrieves the adequate information from these sources and thereafter returns a response to the user. (Hjalmarsson, 2002, p. 13)

2.2.1.5 Language generator/ response generator

When the requested information is retrieved from the external source it needs to be processed in order to be delivered to the end-user. This happens in the language generation step, which means that the message’s (information’s) structure and its’ syntactic structure are determined. One can choose between two alternatives, simple pre-defined templates or complex language generation. In the pre-defined templates the answer is pre-determined and can be used several times with a slight difference in the presentation. This approach is less flexible and sophisticated than the language generation. Language generators plans and structure the message and chooses the most appropriate responding method and the most adequate structure of data presentation before giving a final answer. (Hjalmarsson, 2002, p. 13; McTear, 2004, p. 99)

2.2.1.6 Text-to-speech synthesis

In this step the retrieved response, gained from the earlier step, is in text format and therefore needs to be converted into speech. It is done in this step, where text is converted to speech through speech synthesis. TTS’s are used in unconstrained speech. The first step in the conversion is a text analysis of the input data. The analysis produces a linguistic representation of the text in the form of phonemes. From there the phonemes are generated prosodically, meaning that a description of their rhythm and intonation for example is made available. In the second step, the phonemes are generated into speech (speech generation), which produces a waveform of the speech. This is done by utilizing speech synthesizing methods such as concatenate synthesis. (Hjalmarsson, 2002, p. 13; McTear, 2004, p.102-103)
2.3 Synthesizer technologies

Speech synthesizers are computer systems that produce artificial human speech. The system either converts text to speech (in TTS) or render symbolic linguistic representation (phonetic transcription) into speech. Synthesizers differ in size of stored speech units depending on the purpose and usage domain of the system. For large output ranges the system is normally based on string phones and diphones. The disadvantage of such system is that the speech becomes unclear. For high-quality outputs the system requires that entire words or sentences are stored. In this case the system is usually constrained to a specific usage domain. A good synthesizer is expected to resemble the human voice and also having the ability to make it-self understood. In other words the system needs to possess the qualities of sounding natural and being intelligible. Today there exist several synthesizing methods that can be implemented to synthesize speech. The method used depends on the purpose and the goal of the system. The technologies that are used in the field of speech communication are:

1. Concatenate synthesis, which in turn has three sub-types, unit selection, diphone synthesis and domain specific synthesis.
2. Formant synthesis
3. Hidden Markov Model (HMM)
4. Articulatory synthesis
5. Sinewave synthesis

In the following sub sections the reader’s attention will be turned towards the first three methods listed above since these are the most frequently used today.

2.3.1 Concatenate synthesis

Concatenation is a technology that is based on putting together segments of recorded speech. Until this day this method is the most suitable for giving the most natural and the most intelligible outputs. The drawbacks are that the system becomes restricted to one speaker (one voice) and requires a large memory capacity (Talteknologikursen, 2007). As mentioned earlier, concatenate synthesis is divided into sub-types (US Patent 6665641);

1. Unit selection
2. Diphone synthesis
3. Domain-specific synthesis

Unit selection gives the most natural sounding synthesis due to the minimal amount of added digital signal processing (DSP). In unit selection utterances are segmented into small fragments. Thus larger databases of recorded speech are required to give the most natural sound (Black, A. W., 2002). One of the biggest challenges is to decide which size of the unit segmentation is most suitable for the current system. This applies for all types of concatenate synthesis.

Diphone synthesis, in opposite to unit selection, uses minimal storage memory. The main idea behind this method is to gather all the diphones occurring in a language in a database, which can be used in later concatenation. Diphones are transitions between adjacent phones and this eases the concatenation point of two units and moreover, no second thought is given to the rules of co-articulation. However, diphone synthesis sounds less natural in comparison to unit selection. (Black, A. W., 2000)
The domain-specific synthesis is, as the name implies, restricted to a specific domain. In addition, the concatenation is done between pre-recorded words and phrases to produce utterances. Systems implemented using this method can sound very natural. However, it can only synthesize the combination of words and phrases that it is pre-programmed for. (Talteknologikursen, 2007)

The drawbacks of concatenate synthesis are discontinuities in concatenation points, high memory capacities (especially for long concatenation units). The method is also time-consuming (data collection and labeling of speech samples). The first problem listed can be handled to by using diphone or PSOLA (Pitch Synchronous Overlap Add) methods for smoothing the signal. (Lemmetty, S., 1999; Talteknologikursen, 2007)

2.3.2 Formant synthesis

The rule based synthesis (formant synthesis) uses acoustic models to create synthesized speech outputs. In order to produce the desired utterances the input parameters, such as fundamental frequency, formant frequency, amplitude voice and noise, are varied. For intelligible speech at least three formants are required and for high quality speech up to five formants are needed. Formant synthesis has two basic structures;

- Parallel and

![Parallel Formant Synthesizer Diagram](https://example.com/parallel_synthesizer.png)

\(A_1 \quad F_1 \quad BW_1\)
\(A_2 \quad F_2 \quad BW_2\)
\(A_3 \quad F_3 \quad BW_3\)

Excitation → Speech

(Lemmetty, S., 1999)

*Figure 2.2: A parallel formant synthesizer (where F stands for frequency and BW stands for bandwidth).*

- Cascade

![Cascade Formant Synthesizer Diagram](https://example.com/cascade_synthesizer.png)

Excitation → Gain → Speech

\(F_1 \quad F_2 \quad F_3\)

(Lemmetty, S., 1999)

*Figure 2.3: A cascade formant synthesizer.*

Parallel formant synthesis is better than cascade formant synthesis when it comes to nasal sounds, fricatives and stop-consonants. The cascade synthesizer, on the other hand, is better when dealing with non-nasal sounds. It’s also easier to implement due to its’ simple structure and few input parameters. The best synthesis is given when both methods are combined since they complement one another (Lemmetty, S., 1999). The formant synthesis is the most used
method despite the robotic-sounding. The reason for that is that formant synthesis is able to produce intelligible speech at high levels without disruptions (or technical bumps) like in concatenated speech. Further, it requires neither large memories nor microprocessors due to its small size and the lack of a database. Formant synthesizers have the ability to convey different emotions and tones since the system has control over the output speech and can therefore vary the prosody and intonation of the output data (Lemmetty, S., 1999).

2.3.3 Hidden Markov Model (HMM)

HMM is a finite set of states where each state is associated with a probability distribution. Transition probabilities are the probability sets for transitions between states (Masuki, T. 2002; Warakagoda, 1996). An outcome or observation can be generated in a particular state depending on the probability distribution. Only the outcome is visible to a user not the state. In this model the frequency spectrum, the fundamental frequency and the duration of the speech are modeled simultaneously by HMMs. The speech waveforms are generated from HMMs them-selves based on the maximum likelihood criterion (Masuki, T. 2002). HMM synthesis doesn’t require recordings of large databases to produce a variety of styles (Zen, H. et. al.). The advantages of the method are that modifications can easily be made to the voice, speaking style and emotions by transforming HMM parameters using interpolation, adaptation, eigenvoice, multi-regression etc. HMMs can be trained automatically and are simple and computationally feasible to use.

The HMM speech synthesis consists of two states (Zen, H. et. al);

1. The training stage and
2. The synthesis stage

In the first state, mel-cepstral coefficients (a representation of the short-term power spectrum of a sound) are obtained from speech databases by mel-cepstral analysis. Dynamic features, i.e. delta and delta-delta mel coefficients, are calculated from mel-cepstral coefficients and their deltas and delta-deltas. In the second state, the text which is going to be synthesized is transformed into a phoneme sequence. A HMM sentence, which represent the whole text to be synthesized, is created by concatenating phoneme HMMs. From this a speech parameter generation for HMM using MLSA (Mel Log Spectral Approximation) speech is synthesized from the generated mel-cepstral coefficients. HMMs are trained by using several phonetically balanced sentences uttered by a speaker. (Zen, H. et. al ; Gold & Morgan, 2000, p. 341 ff; Warakagoda, 1996).

2.3.4 Synthesis of disfluent speech

Disfluent speech as explained in section 2.1 is a phenomenon that interrupts the flow of speech and that also occur very frequently in the daily conversation that takes place between two or more human beings. In order to build systems that can imitate the conversational style used by humans these need primarily to comprehend the structure and functionality of the interruptions. In order to do so, researchers have done some studies where they synthesize disfluencies using different methods. Since disfluencies are hard to comprehend just from studying their transcription, a more advisable approach is using the prosodic features and the voice quality of the speech data (Campbell, 2007, p. 22). Non-verbal speech is a way of signaling feelings, acknowledgments and planned actions, which depending on the prosody and voice quality can give the same utterance different meanings. In his study Campbell
(2007, p. 24-25) used unit selection for the synthesis and Tcl/Tk-Snack (14) to extract the acoustic and prosodic features of the non-verbal speech. For the lateral part it means that a vector of 14 values; 5 for pitch, 4 for signal processing, 1 for duration and 4 for spectral characteristic, were used to represent the waveform of speech. The interaction of these four components is considered of great importance while signaling affection and social relationships.

Another experiment conducted by (Carlson, Gustafson and Strangert, 2006, p.22-23) is based on invoking hesitations by using manipulated synthetic stimulus with respect to duration, fundamental frequency and also with the presence versus the absence of creaky voice. The results showed that increased durations, achieved by the combination of retardation and pause, are very good indicative cues on perceiving hesitations. Variations in fundamental frequency and creaky voice due also work as good cues but they are rather of supporting character than primary.

Adell et al. also did an investigation of the most important features for synthesis of conversational speech and, as Campbell (2006), they came to the conclusion that when synthesizing disfluent speech you should study the acoustic features of the voice rather than the transcription features of the speech. The study concentrates on two main disfluency types; filled pauses and repetitions. Hence, emphasizing their prosodic features. According to the investigation done by Adell et. al., and also other work done earlier in speech production, it was confirmed that filled pauses have lower pitch value than their context and that their duration extends from 100 ms to 200 ms. In repetitions, the reparandum and the repair are the same under regular circumstances, in other situations however, it was conducted that the pitch value of both the reparandum and the repair are the same whereas the duration of the reparandum is higher than both the usual value and the repair. The duration of the repair is lower than the usual value. So the synthesis of the disfluent speech was done using these rules before the unit selection and the concatenation steps. The results were somewhat satisfactory though some repetitions were comprehended as synthesis error. The overall quality of the system did decrease, meaning reducing the intelligibility of the system, though a increase in naturalness was observed (Adell et. al.).

Filled pauses (Sundaram et al., 2003) are used differently depending on the occasions and the speaker. However, cues such as “uhs” and “ums” are used to indicate minor and major delays respectively and they are also used during formulation problems. These utterances are also used by the listener to hold the floor. Breathing pauses, on the other hand, are used at breaks in spontaneous speech. Sundaram et al. (2003) states also that breathing pauses are of a non-grammatical character in hesitant speech in comparison to fluent speech. Thus, one can control the fluency of synthesized speech by using these parameters.
2.4 Transcription alphabets

Transcription in speech synthesis means that recorded speech is represented with alphabets, thus, when synthesized the material of the synthetic voice will sound as equal as possible to the original source. In this section three different alphabets will be discussed, namely

- International Phonetic Alphabet (IPA)
- Speech Assessment Methods Phonetic Alphabet (SAMPA)
- The Swedish Transcription Alphabet (STA)

2.4.1 IPA

IPA is an international standard representation of spoken language through phonetic notation that is revised by the International Phonetic Association (IPA, 2005). The IPA is based on the Latin alphabet and it only represents distinctive qualities of speech. Further qualities are represented in the Extensions of the IPA. IPA provides one symbol for each distinctive sound (IPA, 2005; Wikipedia).

2.4.2 SAMPA

SAMPA is another kind of phonetic notation though there’s a lot of resemblance to the IPA. SAMPA is a phoneme notation and therefore requires a different table of alphabets for each language, whereas, IPA is an international table. SAMPA is applicable in many countries but it is still not universal (international). It is applicable for the Swedish language among others. The difference from IPA is that SAMPA letters do not require any spaces between successive symbols and also uses notations that can be typed on a regular keyboard (UCL, 2005).

2.4.3 STA

STA is a phonetic notation for the representation of Swedish spoken language. The Swedish letters Å, Ä and Ö are also included in this alphabet. Further, distinction is made between plosive and non-plosive phases as well as aspirational and non-aspirational phases. Also note that non-verbal sounds and grunts are possible to transcribe with different alphabetical compositions. For example, truncations are marked with a Q after the last pronounced alphabet in the utterance. Inhalingation are represented as INH and conversely, exhalingation are represented as EXH. All sounds are represented by either a single or a composition of Swedish letters.

(For further reading about how different sounds are represented with the different transcription alphabets see section 9.3 (IPA chart, 2005; Salvi, G.)
2.5 Methods for data collection

In year 2001 Shriberg (p. 154) carried out a study on the occurrence of disfluencies in natural speech. Three different domains were tested;

- Human-human free conversation (Switchboard)
- Human-human air travel dialogue (AMEX)
- Human-computer air travel dialogues (ATIS)

According to the collected corpus the analysis showed that the human-human free conversation gave rise to the most amounts of disfluencies.

Other studies done used the method talk-in-interaction (Curl, Local & Walker, 2003, p. 5). According to the authors, this method is the most suitable for analyzing disfluencies in spontaneous speech. They argue that the interlocutors are not restricted to any specific domain which gives broader and larger analytic resources. Secondly, talk-in-interaction is the main tool in social communication and can therefore give us an insight into the world of “everyday language”. A problem with this method is that constructed data may not represent the actions produced (during the performance of the experiment) accurately and therefore not giving solid results.

Kathryn Bock (1996, p. 401) quoted Golden-Eisler (1968) saying, in short, that for research in the field of spontaneous speech it is preferable if the experiment is restricted to speech which is recorded during controlled conditions. Further, Bock (1996, p. 407) discusses the benefits and fallbacks of controlled and uncontrolled input and output. When using controlled conditions in a study a researcher risks to manipulate the data to his/her benefit. On the other hand, if the experiment is performed during uncontrolled conditions there is a risk that the conversation goes out of hand and leads to lots of unnecessary material. On the same page the author brings to light two main methods that are used in speech production;

- Priming and interference techniques, which give rise to both normal speech and speech error
- Manipulation of messages and the pragmatic context in which the messages are produced

Later on in the report Bock (1996, p. 409) mentions that using methods that are based on competition, for example between two interlocutors, increases the rate of errors due to the stress and pressure the candidates are put through. This leads to interruptions in speech. Another efficient method is interrupting the subject’s primary task by distracting events which will take the interlocutor by surprise and hence interrupt the flow of speech (Bock, p. 413). Other oppositions would be to try to influence the candidate’s perspective while keeping the original content of the message (Bock, p. 416). According to Bock this is the most efficient method for deriving interruptions in spontaneous speech. One can also make the candidates recall a word or picture mentioned or showed a bit earlier in the discussion and hence evoke a disturbance in the speech caused by the reason that they need to think back for a few seconds before giving an answer. The aim with this procedure is to make the speaker hesitate, pause, repeat etc. some utterances until they’ve structured their next sentence (Bock, 1977; Bock & Brewer, 1974; Bock and Irwin, 1980; Bock and Warren, 1985; Kelly, Bock and Keil, 1986; McDonald, Bock and Kelly, 1993) referred to in (Bock, 1996, p.413-416). The reason why this method can give good results for analyzing disfluencies in speech is that humans have
difficulties in remembering what they have heard exactly. What is worse is that they have even bigger problems in recalling what just have been said to them in overall (Bock, 1996, p. 416; Potter & Lombardi, 1990; Jarvella, 1971).

2.5.1 Blaue, Eileen & Kay

Blaue, Eileen and Kay did an experiment where in the first part they had 61 students that listened to monologues under three conditions;

1. The monologues were spoken in normal speed
2. The monologues were spoken with 3-seconds pauses that were inserted every 23 words, on average.
3. The monologues were spoken with filled pauses, hesitations.

The results showed that the subjects comprehended the filled pauses significantly better than the monologues with normal speed. And the blank pauses in condition two were slightly harder to understand than condition 3. In the second part of the experiment Blaue, Eileen and Kay divided the candidates in four groups, whereas 3 of them listened to the monologues in section 1 and the fourth group listened to a mechanically slowed version of the original monologues. Even here, the filled pauses gained more points than the other three versions. In overall, one can say that the insertion of hesitation markers eases the comprehension of the content of the speech more than the other versions.
2.6 DEAL

DEAL is the name of a game invented to support foreign students in their language learning and is in reality a complement to Ville, a virtual language tutor. The goal of the game is to improve the students’ conversational skills through a bargain between the user and the ECA (Embodied Conversational Agent), animated talking agent. The game is constructed so that a role-play takes place between the user and the ECA in a flea trade-market. The ECA represent the shopkeeper and the user is the customer. To keep the conversation in flow the customer aim to buy the items using only a certain amount of money and at the same time the shopkeeper tries to negotiate higher prices for his objects.
(For further reading see Klarenfjord, 2008, p. 6-9)
3 Method

This section consists of two main parts, data collection and listening test. The first part will introduce the reader to the scenario design, to the recording process and to the synthetic generation. The process of synthetic generation includes segmentation of corpus (annotation), transcription and synthesis. In the second part of this chapter the reader is given a chance to become familiar with the performance of the listening test.

3.1 Data collection

In order to build a dialogue system that can interact with humans in the same way as humans interact with each other, meaning speaking spontaneously, there is a need for studying different recording methods that will give rise to credible and applicable data collection for synthetic generation. Since the goal is to have a spontaneous conversation with a machine it is appropriate to use recordings of human-human conversation. A second approach is to read pre-written sentences where different disfluencies are simulated. In order to reach the goals of this thesis three hypotheses were used as guidelines in designing the recording sessions;

1. Synthesis generated for use in spoken dialogue systems sounds better if they are based on speech corpora provoked in a dialogue situation rather than in laboratory read.
2. Synthesis of disfluent speech can be generated by using two methods; speech which is recorded in a dialogue or by played disfluencies (laboratory read).
3. Disfluencies can be elicited by putting up a role-play that is performed by two individuals, whom are restricted to a specific domain. The elicitation is performed by a third person that eavesdrops on the conversation that takes place between the two role-players and sends them instructions.

3.1.1 Scenario design

This experiment was carried out in Swedish. All participants are native Swedes and have Swedish as their native language. The process of designing the scenario took a few weeks to finish since many small details needed to be decided and counted for before the actual script was set. The first part of the scenario design was to decide how to carry out the experiment. That is, what would be the most suitable set-up for the data collection; a human-human conversation, a picture description, human-computer conversation, a monologue or an interview? The goal was to record spontaneous speech which can only be provoked when two human beings are speaking to each other under less constrained conditions and therefore it was decided to use a human-human conversation. At the same time a face-to-face conversation was not so desired in order to keep the illusion of speaking to a machine, which lead to the separate room scene setup and SynFace (Karlsson, I., 2001-2004; Edlund and Beskow). The next step was to agree upon a discussion theme for the recordings. Not long before this experiment became a high topic another project, called DEAL, was carried out by researchers at the department of TMH (Speech, Music and Hearing) at the Royal Institute of Technology. DEAL, which can be read about in section 2.6, is a game that takes place in a trade market and is primarily constructed to support exchange students in their language learning. The recording results from the DEAL-project were good and for that reason colleagues at TMH recommended that a similar interface would be suitable in this case also. Therefore the theme for this project became on-line selling products, where the interlocutors interacted with each other over an audio channel called Skype (Thomann; Wikipedia). The
next step was to specify what type of disfluencies to provoke and how these can be provoked. The disfluency types that were decided to study are;

- Hesitations
- Pauses
- Repairs
- Repetitions
- Truncations (cutoffs)

The reason for why this specific categorization is used is based on two factors; the most frequent occurring disfluencies and the best descriptive name of a phenomenon. Clues for three of the disfluency types listed above are presented in table 3.1 below. These were used later on in the design of the recordings. Repairs and repetitions can be provoked in parallel with hesitations, pauses and truncations, therefore, no specific column is dedicated to these.

<table>
<thead>
<tr>
<th>When do hesitations occur?</th>
<th>When do pauses occur?</th>
<th>When do truncations occur?</th>
</tr>
</thead>
<tbody>
<tr>
<td>• A person needs to decide between two equally valuable persons or material things</td>
<td>• When a person is given a tricky or somewhat odd question that requires that the person thinks before he/she gives his/her answer</td>
<td>• When two or more persons are having a discussion about a subject that is emotionally loaded for the parties involved</td>
</tr>
<tr>
<td>• The person is uncertain of what he/she is saying because he/she does not remember the story in detail.</td>
<td>• When the person is choked by a comment and needs a minute to recover before responding</td>
<td>• When correcting the person standing in front of you</td>
</tr>
<tr>
<td>• The person knows that an action is wrong but if he/she does not perform it the outcome will be even worse.</td>
<td>• When a person discover that what he/she just said is not fully true, so he/she needs to correct him-/her-self.</td>
<td>• When a person is impatient, defensive or annoyed he/she has the tendency to interrupt the other person standing in front of him/her</td>
</tr>
<tr>
<td>• A person is given a question that can have several answers and depending on the motivation they all can be classified as right</td>
<td>• When given a difficult question where the answer is not obvious</td>
<td></td>
</tr>
<tr>
<td>• A person is given a question knowing that if he/she gives an honest answer he/she is in trouble. And by giving a dishonest answer he/she would be lying.</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

The following step was to decide what type of instructions would result in provoking the different disfluency types. As my supervisor Joakim Gustafson once wrote to me, natural dialogue phenomena can be called upon through giving the interlocutors;

1. A persona description of the role they will be playing
2. Steady instructions related to the task they will perform
3. Dynamic instructions

- Graphical instructions (pictures that can be manipulated in some way)
- Textual instructions
  - Emotional
  - Task-related
  - Keeping the dialog in flow

*Emotional* instructions tell the receiver to act emotional (e.g. act grumpy); *Task-related* instructions require the receiver to initiate a certain sub-task (e.g. buy a car); and *Dialogue flow related* instructions tell the receiver to change his way of speaking, (e.g. speak fast without pauses).

For the first recording session all types mentioned above were combined together to avoid silence or short dialog flows. Hence, not all persons can act naturally when limited to only one of the scenario types. However, in recording two the steady instructions were excluded since they were found to be unnecessary and did not contribute with anything when combined with the dynamic instructions. The next step was to define each and every type of the dynamic instructions in relation to the different disfluencies. While studying table 3.2, it becomes clear that one disfluency type can be provoked in several ways. This table was constructed using real-life experiences of human-human conversations. That is, how does a person respond to the question;

What is 3-(3x4)? (Task related)

The typical answer would start with the utterance ahhh, which belongs to the disfluency type called hesitation. Similar questions were used throughout the construction of table 3.2.

<table>
<thead>
<tr>
<th>Disfluency type</th>
<th>Graphical</th>
<th>Emotional</th>
<th>Task related</th>
<th>Dialogue flow</th>
</tr>
</thead>
<tbody>
<tr>
<td>Hesitation (S*)</td>
<td>Change picture in mid speech (S) Show an ambiguous picture (S)</td>
<td>Act as if you won the lottery (B) Be cheap (B)</td>
<td>Buy red car (B) at the same time as Sell blue car (S)</td>
<td>Interrupt when it’s silent (B)</td>
</tr>
<tr>
<td>Hesitation (B*)</td>
<td>Change picture in mid speech (B) Show an ambiguous picture (B)</td>
<td>Be hysterical (S) Say contradictory things-change point of view regularly (S)</td>
<td>Buy red car (B) at the same time as Sell blue car (S)</td>
<td>Interrupt when it’s silent (S)</td>
</tr>
<tr>
<td>Stumble (S)</td>
<td>Change picture in mid speech (S) Show an ambiguous picture (S)</td>
<td>Laugh loud and totally unexpected (B)</td>
<td>Reduce the price with 10% (S) Compute the profit (S)</td>
<td>Say what continuously (B)</td>
</tr>
<tr>
<td>Stumble (B)</td>
<td>Change picture in mid speech (B) Show an ambiguous picture (B)</td>
<td>Laugh loud and totally unexpected (S)</td>
<td>Quandar (B) Add 25 % of tax to the price (B)</td>
<td>Say what continuously (S)</td>
</tr>
<tr>
<td>Pause (S)</td>
<td>Change picture in mid speech (S)</td>
<td>Be rude (B)</td>
<td>Reduce the price with 10% (S) Compute the profit (S)</td>
<td>Speak slowly (S) Pause (S)</td>
</tr>
<tr>
<td>Pause (B)</td>
<td>Change picture in mid speech (B)</td>
<td>Be rude (S)</td>
<td>Quandar (B) Add 25 % of tax to the price (B)</td>
<td>Speak slowly (B) Pause (B)</td>
</tr>
<tr>
<td>Truncation (S)</td>
<td>Change picture in mid speech (S)</td>
<td>Start laughing loud in mid-speech (B)</td>
<td>Give a price (B) at the same time as Give a price (S)</td>
<td>Speak as long as there’s someone else speaking (B)</td>
</tr>
<tr>
<td>Truncation (B)</td>
<td>Change picture in mid</td>
<td>Start laughing loud in</td>
<td>Give a price (S) at</td>
<td>Speak as long as</td>
</tr>
</tbody>
</table>

Table 3.2: Dynamic instructions and the corresponding interruptions.
<table>
<thead>
<tr>
<th>Speech (B)</th>
<th>Mid-Speech (B)</th>
<th>The same time as</th>
<th>There's someone else speaking (S)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Repairs (S)</td>
<td>Change picture in mid speech (S)</td>
<td>Use synonyms (S)</td>
<td>Mix up the colors (S)</td>
</tr>
<tr>
<td>Repairs (B)</td>
<td>Change picture in mid speech (B)</td>
<td>Use synonyms (B)</td>
<td>Mix up the names of the products (B)</td>
</tr>
<tr>
<td>Irritation (S)</td>
<td>Show an ambiguous picture (S)</td>
<td>Show dissatisfaction (B)</td>
<td>Ask why frequently (B)</td>
</tr>
<tr>
<td>Irritation (B)</td>
<td>Show an ambiguous picture (B)</td>
<td>Show dissatisfaction (S)</td>
<td>Buy red car (B) at the same time as sell blue car (S)</td>
</tr>
<tr>
<td>Uncertainty (S)</td>
<td>Show an ambiguous picture (S)</td>
<td>Show dissatisfaction (B)</td>
<td>Point out defects in the products (B)</td>
</tr>
<tr>
<td>Uncertainty (B)</td>
<td>Show an ambiguous picture (B)</td>
<td>Show dissatisfaction (S)</td>
<td>Give unreasonable prices (S)</td>
</tr>
<tr>
<td>Task (S)</td>
<td>Show a picture (S)</td>
<td>Talk uninterruptedly about the product’s qualities (S)</td>
<td>Sell red car (S)</td>
</tr>
<tr>
<td>Task (B)</td>
<td>Show a picture (B)</td>
<td>Talk uninterruptedly about the qualities you seek in the product you which to purchase (B)</td>
<td>Buy blue car (B)</td>
</tr>
<tr>
<td>Repetition (S)</td>
<td>Show an ambiguous picture</td>
<td>Show dissatisfaction (B)</td>
<td>Sell blue car at the same time as you know that red car is best (S)</td>
</tr>
<tr>
<td>Repetition (B)</td>
<td>Show an ambiguous picture</td>
<td>You have a decision dilemma (B)</td>
<td>You want the red car but you hesitate before you make up your mind (B)</td>
</tr>
</tbody>
</table>

*S stands for seller and B stands for buyer

The last step in the scenario design was to decide what is and respectively what is not typical for a sell-and-buy situation;

- **Typical situations**
  - ✓ Choose between several products
  - ✓ Being indecisive
  - ✓ Choose a specific product
  - ✓ Have change of heart due to a high product price
  - ✓ The product is sold out
  - ✓ Being obliging
- **Atypical situations**
  - ✓ Unreasonable high prices
  - ✓ Change the price suddenly
  - ✓ The seller acts like a psychopath
  - ✓ The seller is rude (it happens but it's not that common)
  - ✓ Being unreasonably happy or sad because a product is outsold
The seller tries to convince a customer to not by a product

A complete version of the scenario with instructions and the expected outcomes are available in appendix A (the text is in Swedish).

3.1.2 Recordings

This project is based on two recording methods in order to test the first and second hypothesis of this study; a human-human conversation and read sentences. In the first setup the scenario resembles the main idea behind the game DEAL. A seller and a buyer are interacting with each other in a trade-market. In addition, a so called human-wizard controlled the flow of the conversation, thus, making the seller and buyer function as puppets (testing the third hypothesis). The wizard followed a pre-written manuscript of the instructions. The candidates were handed a persona description before the experiment was started and also some steady instructions (for recording 1 only). The seller was a male and the buyer was a female in both recordings one and two. In recording 3 only the male candidate participated, since it was a monologue. All three participants active in the recordings, except for the wizard, know each other and have worked together earlier.

Three recordings were made and each one of them had a recording time of about 40 minutes. In total, a corpus of about 2 hours was gathered and evaluated for further investigation. The recorded material was segmented in all three occasions which gave approximately 250 sentences. From these only 24 speech segments were chosen for the listening test.

3.1.2.1 Recording 1

The first recording session was based on a controlled human-human dialog. Two participants were given the instructions to act as a shop-keeper and a customer respectively. The shop-keeper’s personality was enrolled by a male research engineer. The opponent was a female PhD student. Both worked at the department of Speech, Music and Hearing at KTH at the time this experiment was executed. Each candidate was handed a persona description minutes before the actual start-up of the recording. In addition, they were given steady instructions to follow as the role-play went along (both the persona description and the steady instructions are available in appendix B and C respectively). The participants were seated in separate rooms with a prompter and a headset. The session was recorded using Wavesurfer. Hence, the recordings of the speech data were done using a separate channel for each participant. The files were later synchronized, keeping in mind that a 200 ms delay in each direction was introduced when using SynFace. The computer screens of both interlocutors were connected to the wizard’s computer, which was located in a third room. Like the interlocutors, the wizard had also access to headphones in order to eavesdrop on the conversation of the two participants. The interface of the seller and buyer included a SynFace (an animated head) representation of the opponent to the left and a product representation to the right. Below these, the instructions sent by the wizard scrolled in from right to left. The wizard communicated with the participants through an automated control interface. That is, all instructions and products were preprogrammed and the wizard only needed to choose (not type the instruction by hand at present time) the type of instruction he wished to send to the participants. This service was used mostly to spear time and avoid typing mistakes. The manuscript that was used for this recording is available in appendix D. Figures 3.1-3.2 on the following pages give an overview of the respective interfaces.
The wizard’s control interface with a copy of the interlocutors’ interfaces at the top.

The wizard had access to both interlocutors’ interfaces at the top; the seller to the left and the buyer to the right (figure 3.1). At the bottom, the different types of dynamic instructions were made available for the wizard to use during the recording session. The bottom-left sent instructions to the seller and the bottom-right sent instructions to the buyer.

a) The seller’s interface with the buyer’s lip synchronized avatar
3.1.2.2 Recording 2

Recording session two followed the same structure and scene setup as recording session one. The main difference was that the participants and instructions had changed a bit. The shopkeeper was played by the same research engineer from session one, but the opponent was a female Master Thesis Student. Moreover, some improvements were done to the manuscript after the analysis of the collected data from session one (for more details about the structure of manuscript 2 see appendix E). When compared to manuscript one it can be seen that some instructions have been either modified or removed entirely. Hence, some of the instructions did not evoke the requested reactions or disfluencies and were therefore removed. Other instructions were considered to have a potential but due to their ambiguous content they were hard to follow on the dot and were therefore shortened or done more plain. See row 5 and 54 in the manuscripts for clarification (are ). The former is a modification and the second example is removed entirely in manuscript two since it did not give the expected outcome.

Row 5

K: Missnöjd med alla → K: Du är inte riktigt nöjd med förslagen som försäljaren kommer med

Row 54 (in manuscript 1)

F: Prata snabbt och osammanhängande

Another modification done to session two was that the steady instructions were removed since they were considered hard to follow and at the same time keep track of the dynamic instructions given by the wizard.
3.1.2.3 Recording 3

The third recording differed entirely from the first two recordings. The reason was to collect speech data from two different recording methods in order to test the hypothesis of this project. Hence, which method gives the most natural synthesized disfluencies; the spontaneous dialogue or the played dialogue?

The male participant from the earlier recordings was the only active candidate in this session. A few speech segments including disfluencies were chosen from the male’s first two recordings. The interruptions were identified and then put into new contexts. The candidate was instructed to read the original sentences, then the modified sentences with the disfluencies and last the modified sentences without the disfluencies. The data used for the last recording is available in appendix G. Also this session is recorded using wavesurfer.

3.1.3 Transcription

When the recorded material from session one and two was synchronized, specific parts were annotated and transcribed with HigginsAnnotationTool (Klarenfjord, 2008, p.27). The transcription of the chosen speech segments were done by hand. With this word transcription step the aim was to represent the spoken words exactly as they sound in written form for later phonetic transcription and diphone synthesis. Figure 3.3, on the next page, shows an example of the word transcription in HigginsAnnotationTool taken from the experiment. The speech signal highlighted with red corresponds to the text in the second box (also highlighted with red).

![Figure 3.3: HigginsAnnotationTool with one channel per speaker.](image)

After the word transcription of the speech, the sound segments were phonetically transcribed and aligned to build small diphone voices with the EXPROS tool (Gustafson and Edlund,
2008) and manually corrected in WaveSurfer using STA. The reason for using STA (The Swedish Transcription Alphabet) as the phonetic transcription language is due to the fact that the experiment was carried out in Swedish and therefore this technique was found more adequate than the other alphabets mentioned in this report. The voices produced with the EXPROS tool contained fillers, truncated phonemes and audible breathing. Non-verbal sounds, such as laughter, coughs and exhalations and inhalations are also transcribed (see table 3.4).

Table 3.4: STA for non-verbal sounds

<table>
<thead>
<tr>
<th>Type of disfluency</th>
<th>STA representation</th>
</tr>
</thead>
<tbody>
<tr>
<td>Hesitation</td>
<td>ÖHH, EHH, MMM, ÖHM, EHM, AHH</td>
</tr>
<tr>
<td>Laughter</td>
<td>SKRATT</td>
</tr>
<tr>
<td>Cling with tongue</td>
<td>SMACK</td>
</tr>
<tr>
<td>Exhalation</td>
<td>EXH</td>
</tr>
<tr>
<td>Inhalation</td>
<td>INH</td>
</tr>
<tr>
<td>Hawking</td>
<td>HARKL</td>
</tr>
<tr>
<td>Truncated phonemes</td>
<td>Were marked with Q</td>
</tr>
</tbody>
</table>

3.1.4 Synthetic generation

In order to test the hypothesis on which data to use when building speech synthesizers for use in spoken dialogue systems a customized mini voice for the MBROLA diphone synthesizer was created. The diphone voice was generated using the EXPROS tool and the MBROLA diphone voice creation tools (Dutoit et al). 20 utterances from the dialogue collection were selected since they contained the interactional phenomena (disfluencies) this project wish to be able to synthesize. The sentences included disfluencies like hesitations, cutoffs, pauses, corrections and repetitions. In order to verify that the disfluencies could be synthesized in different contexts, a number of new test sentences were constructed that included the same immediate contexts as in the 20 spontaneous conversational utterances. A transcript of the original utterance without the disfluencies was also constructed. Finally, a version of the new test sentence with added tags for disfluencies (ehh, silence) was created. All sentences were read by the original male speaker (the seller). All recordings, both the spontaneous conversational utterances and the three sets of read sentences were processed by the EXPROS tool to generate the data to build a small customized MBROLA voice (called Own) that could be used to synthesize all four sets;

- the original conversational sentences with real disfluencies
- the read version of the same utterances without disfluencies
- the constructed read sentences without disfluencies
- the constructed read sentences with played disfluencies

The disfluent phones in the Own voice were tagged the STA that had been extended with non-verbal sounds (see Table 3.4 above). This meant that extra linguistic sounds like audible breathing could be inserted into the read sentences and that they could be stopped abruptly by inserting a truncated phoneme.

First all sets were re-synthesized with the new Own voice and with a Swedish commercial diphone voice that is trained on clear read speech (henceforth Ingmar). By re-synthesizing the sentences all prosodic features (pitch, duration and loudness) are kept from the original recordings, but the voice quality is modified - either to the Own diphone voice with its artifacts of the signal processing involved in creating a diphone voice, or to the Ingmar voice that apart from these artifacts also has a different speaking voice and it is trained on clear
speech. In order to get a golden standard when evaluating the naturalness and clearness all original recordings were encoded with wideband AMR coding, which is commonly used in cell phones. This encoding also introduces signal processing artifacts, but on a level that has been chosen to be acceptable for human-human communication. These three types of sentences will be referred to with Ingmar, Own and AMR henceforth.

Secondly a number of synthesized utterances were generated, where disfluencies were inserted into the read sentences that did not contain such from the beginning. This was done by modifying the MBROLA input files that had been used to generate the re-synthesized sentences above. The way the disfluencies were inserted differed for the two voices: for Ingmar the pitch, loudness and durational information from the disfluent part of the original utterances were inserted into the MBROLA input files of the read sentence. When the Own voice was used, the new STA labels for disfluent phones were also moved in this process. This meant that the generated sentences kept its read and fluent prosody and voice quality in most of its parts, with conversational or disfluent prosody and voice quality only in the inserted disfluent parts. The disfluent parts were either taken from the original dialogue recordings or from the set of read sentences where the subject was asked to read them with transcribed disfluencies. This set of synthesized sentences will be referred to as real_to_read_sentence respectively played_to_read_sentence henceforth.

Finally, a small set of sentences where the speakers interrupted each other were selected. In half of these the other interlocutor who stole the turn from the male speaker, in the other half he fought back and continued talking after the interruption. A small diphone voice was created from both types of interruptions, which made it possible to change the way a re-synthesis of the utterances went silent (as in abandonment or as when taking back the turn). A set of sentence were synthesized where it simply re-synthesis of the two ways of going silent in their original sentences. In a second synthesis generation the diphones created at abandonment interruptions where used in sentences where the speaker actually took back the turn, and vice versa. These sentences were used in the last part of the listening test and in the result chapter they are referred to as break (abandonment) and pause (taking back the turn).
3.2 Listening test

When all data was collected and the synthetic generation was completed the next step was to test the corpus. The method used was a listening test, where about 16 persons listened to some chosen segments of the recorded speech and thereafter answered a few questions. The answers from the listening test were used to evaluate whether or not the goals of this project were reached.

3.2.1 Method

The purpose of the listening test is to compare the different voices and methods used during the execution of the experiment. This part of the experiment is crucial in order to verify whether or not an outsider can;

1. Hear the occurrence of a disfluency
2. Identify the type of a disfluency
3. Accept the disfluency for being natural (a spontaneous act)

Moreover, the voices Own, Ingmar and AMR are compared as were the dialogue and read recording methods.

The test was based on five parts;

1. Personal questions about the subject
2. Comparison of the different voices
3. Comparison of the different recording methods
4. Cues for turn-taking
5. An overall comparison of the voices and suggestions for future improvements

3.2.2 Selection of subjects

The goal with this project as recalling from earlier sections was to find a recording method that give human-like synthetic disfluencies. At the department of Speech communication at KTH, where this project is performed there are many skilled researchers and PhD students that can participate in the listening test and give a hand-full of constructive feed-back. However, the drawbacks of using people from the department would be that they are too involved and have too much knowledge in this area of work and can therefore not represent the naïve user. For this specific reason the subjects that were chosen for performing this test are all, except for two, persons with no or minimal knowledge about how SDS:s work.

16 subjects were asked to perform the listening test. All participants are Swedish citizens and speak very good Swedish. However many of them have Swedish as a second language. They are all of different ages and have different academic backgrounds. Many have higher education or are in the process of getting one. Others have only a high school graduation diploma. Few are engineer students and hence they are not that familiar with SDS’s. Almost all have former experience of spoken dialogue systems, but not of the advanced kind. That is, they have mostly used answering machines and simple support systems or educational programs. Only two of all 16 subjects are well introduced to the area of communication speech and SDS: s.
3.2.3 Settings/platform

The subjects performed the test one at a time, in a quite silent and distraction free room. The test was performed on a computer with headphones. The objects’ answers were registered in separate logs that are created at the startup of the test. Figure 3.4 gives an example of a typical log used for the listening test. In the first row the answers from part one are registered, such as name, gender, profession etc. The log is divided in 7 columns (all seven columns are not used in the fourth and fifth part since they contain fewer questions);

1. Name of the subject
2. Name of the test
3. Name of the speech file
4. Number of times that the subject listened to the same file
5. Type of identified disfluency/who will continue to speak
6. How easy the disfluency could be heard
7. How natural the disfluency sounded

![Figure 3.4: The log’s interface when the listening test is performed.](image)

Figure 3.5 gives an illustration of the layout of the different parts of the listening test. As recalling, part one only consists of personal questions concerning the subject. In the second part, the subject listens to a speech file and thereafter identifies the type of disfluency and also grade the clearness and naturalness of the speech. A 6-point Likert scale (Likert, 1932) is used in order to assess the acceptability, naturalness and comprehension of disfluent speech with the different synthetic voices. In the third part, the listener is asked to identify as well as compare two different speech files. The purpose is to assess the acceptability, comprehension and naturalness of the different recording methods. The fourth part tests whether or not the cues; pause and break, can work as indicators of who will continue to speak after an interruption. In this study the files are constructed as such that when a pause is introduced, the same speaker will continue to speak after the interruption, whereas when a break is introduced a second speaker will continue to speak. The fifth and last part is just an overall assessment of the different voices used in this study.
a) The first part of the test where the subjects answer a few personal questions.

b) Comparison of the different voices.

c) Comparison of the different recording methods.

d) Cues for turn-taking.

e) An overall comparison of the voices and suggestions for future improvements.

Figure 3.5: The interface of the listening test.
4 Result Analysis

In this chapter all recordings as well as all collected data will be analyzed and evaluated. First, an evaluation by the interlocutors, concerning the recording methods, will be presented and also the observations done during and after the recording sessions. The evaluation will be followed by a result analysis of the listening test.

4.1 Analysis of the recording sessions

Three recordings were done in total. Two were based on human-human conversation and the third was read sentences. How efficient the instructions were and how did the different recording setups work will be given in the following subsections.

4.1.1 Evaluation of recordings 1 and 2

Most instructions in manuscript 1 (Appendix D) did evoke the expected reactions and disfluencies, though there were some that did not give the intended result (see Appendix H). Observations and analysis of the first recording showed that the participants gave no second thought to or forgot about the steady instructions that were handed to them seconds before the start-up of the recording. However, they did try to stick to the persona description given for the role-play. The dynamic instructions that did not work as expected were often either difficult to follow or misunderstood (picture disappearing was interpreted as the object had been bought in comparison to being sold to other customers online). Another observation done is that unexpected instructions or peculiar ones provoke much disfluency;

- Changing the prices during speech or at the exact moment as the seller promises the buyer a price.
- Instructions which makes the interlocutors react strongly
- Give the new instructions while the interlocutors are still speaking
- Give opposite instructions to the interlocutors.

Attention should also be drawn to the fact that the manuscript was not followed on the dot. Some other instructions, not included in the manuscript, were used and some instructions included in the manuscript were excluded.

4.1.1.1 The evaluation of the interlocutors

In both recording one and two the female participants found it harder to improvise than did the male participant. Even so, they had fun recording the dialogue and found it very amusing. The same applies for the male candidate, which was very fond of the setup of the role-play. One fallback thou, mostly commented in the first recording, was that the role-play was too long.

While observing and listening to the recorded data attention was drawn to the fact that in both experiments the male candidate contributed with the highest amount of disfluencies than did the two female participants. This actually proves the statement of Elizabeth Shriberg (see section 2.1.3). There was a desire to do a test were both participants were males to investigate the other statement that implies that two male speakers do more mistakes in natural speech than a male speaking to a female or two females speaking to each other, but unfortunately the time ran out.
4.1.1.2 Evaluation of the instructions

Table 4.1: The distribution of instructions that did provoke disfluent speech.

<table>
<thead>
<tr>
<th>Type of instruction that did provoke an interruption</th>
<th>Recording 1</th>
<th>Recording 2</th>
</tr>
</thead>
<tbody>
<tr>
<td>Task related</td>
<td>13/26 (50.0%)</td>
<td>14/24 (58.3%)</td>
</tr>
<tr>
<td>Emotional</td>
<td>9/14 (64.3%)</td>
<td>5/7 (71.4%)</td>
</tr>
<tr>
<td>Graphical</td>
<td>2/4 (50.0%)</td>
<td>4/5 (80.0%)</td>
</tr>
<tr>
<td>Dialogue flow</td>
<td>12/13 (92.3%)</td>
<td>3/9 (33.3%)</td>
</tr>
<tr>
<td>Total</td>
<td>36/57 (63.2%)</td>
<td>26/45 (57.8%)</td>
</tr>
</tbody>
</table>

The dynamic instructions that gave the most disfluency, in total, were the graphical, emotional and the dialogue flow. Table 4.1 shows that the task related worked approximately 54.0% of the time, the dialogue flow 68.2%, the graphical 66.7% and the emotional 66.7% of the time (the rates are based on the total result from both recordings combined). These numbers are obviously rough estimations and not exact rates. Hence, the data is based on only two recordings. However, keep in mind that the instructions gave the best result when they were of an enforcing and a contradictive character. In recording one both the emotional and the dialogue flow instructions contributed with the most disfluency and that was because they, in most cases, were combined or were contradictive, whereas in recording two, the graphical and the emotional instructions gave the best outcome due to the fact that they were combined with contradictive and enforcing task related instructions (contradictive instructions means that both candidates got new opposite instructions at the same time). The most reoccurring disfluencies were hesitations, pauses, repetitions and truncations. In short when aiming for provoking a high rate of disfluency in spontaneous speech it is recommended, from the observations and the analyzed results, that the different types of instructions should be either combined or when only one type is used that they are contradictive. Further, to be certain that the candidates follow the instructions given they should be of an enforcing kind, otherwise there is a risk that the candidate takeOwn initiatives and skip the given instruction. This phenomenon reoccurred several times during recording two, especially at the end.

In order to give an idea of how well the instructions did work, the dialogues in the current domain were compared to the dialogues recorded during a previous project that had a resembling recording method, except for the controlling Wizard, namely DEAL (Hjalmarsson, 2008). In that study 8 dialogues were recorded and about one third of the turns contained disfluent speech. This indicates, when studying the table above, that the disfluency rate found in this current study, due to the instructions, are higher than in the previous DEAL conversations.

In its’ entity one can state that the instructions did give the result that they were assigned for, however not all instructions were completed instantly. At some occasions the interlocutors waited with completing the assignment until they found it appropriate. The instructions came very handy in keeping the discussion between the two interlocutors in flow and in leading the role-play in the right direction in order to give the inquired outcomes. Another observation that was done during the dialogue recordings was that in order to change the direction of the discussion instantly the instruction given must be of an enforcing kind. By that it is meant that the candidates are not given any option but to follow the instruction. One example taken from session one is that the shop-keeper got the instruction to sell a toy to the customer, but as soon as the interlocutors had agreed upon a price, the price changed and the shop-keeper had to
communicate that to the customer. The shop-keeper him-self was in chock since some of the prices were unreasonably high. This in turn gave rise to many interruptions and stumbles.

4.1.2 Evaluation of recording session 3

In order to record more trustworthy and natural sounding speech, without the help of a dialogue (where the conversational partners get feedback from each other) there is a need of rehearsing the utterances. The interlocutor needs to adapt to the phrases that are to be read and, thus, understand what these phrases want to deliver to the listener. Therefore, a need for a good acting talent is desired in this kind of recording setup. The interlocutor would need to exercise and gain feedback before the actual recording day. Self-made recordings would serve as an aiding tool for the interlocutor towards a better and more natural pronunciation, because the aim is to record segments of speech that reflect the spontaneous speech used in human-humans conversations. The ambition is to provoke phrases that consist of both verbal and non-verbal phrases.

4.1.2.1 The evaluation of the interlocutor

The male participant, that recorded all sentences, found it very difficult to reproduce the content like planned. According to the interlocutor’s personal opinion the dialogue form was much easier and more natural to perform since the conditions were less constrained and, hence, gave the interlocutors almost free hands in choosing their expressions. The male participant preferred to talk freely in controlled environments. It is considered of great importance that the person that will read the sentences have good acting skills to be able to deliver the interruptions naturally. In comparison to the first two recordings the last one was very boring to record according to the male participant.

On the basis of the first evaluation of the recording it was found, in agreement with the participant, that the speech sounded very affected and unnatural. Truncations, repairs and pauses were not expressed correctly. The only sentences that resembled the original dialogues were those that included hesitations.

4.1.3 Summary

According to the comments of the participants in the recordings and to the analysis of the recorded data it is obvious that the dialogue recordings give more natural sounding disfluent speech than read sentences (played disfluencies) as was stated in the hypothesis of this study. Disfluencies in a dialogue are easier to comprehend and it is also easier to distinguish the different disfluency types apart when using a recording method based on a dialogue rather than a monologue. In monologues, in agreement with Blaue et al. (see section 2.5.1), the easiest disfluency type that a listener can comprehend is hesitation.

The third hypothesis of this research is proved, thou the number of recordings is too small to give statistically significant results.
4.2 Result analysis of the listening test

All 16 subjects performed the test separately at different occasions. Each test took on average 35-40 minutes to perform. As recalling from section 3.2 the test has 4 main parts.

4.2.1 Voice quality

The results from part two are presented in figure 4.1 below.

![Figure 4.1: A diagram representation of the results from part 2 in the listening test; comparison of the different synthetic voices.](image)

a) How natural the synthesized disfluencies sound using AMR, Own and Ingmar.

b) A gradation of how easily a disfluency is recognized using AMR, Own and Ingmar.

Figure 4.1a shows that AMR is rated as most natural and that Ingmar is the least. The Own voice is rated as significantly more natural than Ingmar, but also significantly less natural than AMR. When examining the rated naturalness of only the truncations no significant difference was found between AMR and Own (Kruskal Wallis p=0.29). However, the truncations
generated in *Own* are rated significantly more natural than the ones synthesized with *Ingmar* (Kruskal Wallis $p=0.001$). Figure 4.1b shows that all three voices are rated as equally clear in their delivery of the disfluencies. There is no significant difference between the ratings of clearness between any of the voices. Unfortunately, these results are not supported statistically.

In part three the different recording methods are compared to each other. Figure 4.2 show that the real dialogues were perceived as real (natural) in 60 % of the time. The surprising result was that the *played_to_read_sentences* (read sentences with played disfluencies) were graded to sound more natural than the *real_to_read_sentences* (read sentences with real disfluencies).

![Figure 4.2: A diagram representation of the data from part three in the listening test; a comparison of the different recording methods.](image)

The diagram representation in figure 4.3 indicates that when *Own* was interrupted the listener got the impression of that the same speaker would continue to speak after the interruption no matter if the disfluency type was a break or a pause. The difference between the two was minimal. For the voice of *Ingmar*, on the other hand, there was a difference of more than 10 %. When an interruption was introduced using a break more than 50 % expected that the second speaker would continue to speak, whereas when the interruption was introduced using a pause about 60 % answered that the same speaker will continue. Thus leading to the conclusion that *Ingmar* is more intelligible, since in this project a pause was used when the same speaker would continue whereas a break was used when the other speaker would continue.
In the last part of the test most listeners preferred the voice of Own over Ingmar, since it sounded more human-like (see appendix L). Ingmar, though having clean pronunciation and being very understandable, irritated many subjects by having a mechanical sound quality.

### 4.2.2 Identification of disfluency type

In part two and three of the listening test subjects were asked to identify the type of disfluency. The data for the identification is presented in figure 4.4, 4.5 and 4.6 respectively.

Figure 4.4 above shows that when analyzing the identification of hesitations, pauses and truncations no significant difference was found between the different voices except for the significantly higher amount of inserted truncations in Ingmar in comparison with Own (Pearson chi-Square p=0.001).

The number of the different disfluency types identified when AMR was played (figure 4.5) were about the same. Hence, the same goes for Own though the number of hesitations was a bit higher than the others. For Ingmar, the diagram shows that the number of occurred pauses, hesitations and repetitions are the same. The largest deviation is seen in the number of truncations (cutoffs). This value is above 80 and is the highest value amongst all. The lowest
number, in all three voices, is represented by repairs. In figure 4.6, only Own and Ingmar are compared, since these are the two main voices that this study wishes to test. Obviously the difference between the two results is minimal. One cannot say that one voice overruled the other like in figure 4.5.

![Figure 4.5: Identification of the different disfluency types for part 2.](image)

![Figure 4.6: Identification of the different disfluency types for part 3.](image)

### 4.2.3 Who will continue?

As explained in earlier sections, part four of the test was based on identifying who will speak after the interruption. In this section the interruptions were divided into two categories; a pause and a break. A pause could either be introduced by a cutoff or was just a silent pause. A break, on the other hand, had several representations;

- Creaky voice
- Cutoffs
- Exhalations
- Inhalations
- Nothing
Figure 4.7 presents the distribution for these cues. The most distinct feature that the same speaker will continue to speak after an interruption is the cutoff in the break category or none in the pause. This applies for both Own and Ingmar. No concrete conclusion can be drawn from these results concerning what disfluency type is directly related to what speaker.

![Figure 4.7: The type of disfluency that reveals the speaker.](image)

### 4.2.4 Summary

The results in section 4.2.1 indicates that independently of who the speaker is the occurrence of a speech disfluency is clearly perceived, however, how natural these interruptions sound differ depending on the speaker. Further, next to the real_dialogues the played_to_read_sentences sounds more natural than the played_sentences and the real_to_read_sentences, according to the analyzed data. Moreover, in figure 4.8 below, listeners identified some of the disfluencies as repairs despite of the fact that no repairs were included in the used files for the listening test. Attention should also be drawn to the fact that not all interruptions were identified correctly all the time. Pauses and repetitions were the two types that were mostly confused with other disfluency types.
In conclusion, the hypotheses that are stated in the introduction are proved. However, due to the restricted amount of data the differences found between the different recording methods and the different synthetic voices are too small to give statistically significant results in all cases. In part two only the difference in naturalness between AMR and Own and also between Own and Ingmar are statistically significant according to Kruskal Wallis Assym significance (p=0.000). For clearness the results were p=0.684 for AMR an Own and p=0.425 for Own and Ingmar. In part three the results for whether the disfluency was played or real were significant for Ingmar but not for Own while using Pearson chi square. The results from the forth part in the listening test were insignificant for figure 4.3 when using Pearson chi square (p=0.15 for Ingmar and p=0.23 for Own) due to that only one cue was used for the turn-yielding and that some interruptions were semantically incomplete (Hjalmarsson, figure 5). In figure 4.7, however, the results are more reliable and are statistically significant while using Pearson chi square (p=0.023).
5 Discussion

The listening test in this experiment was of crucial importance in order to state how the different recording methods and synthetic voices were appreciated by the naive user. Due to a small number of dialogue examples, the small number of test subjects in the listening test and the complexity of their task it became hard to gain significant results that could verify the hypothesis of this study. However, there are some significant results: The disfluencies generated with the dialogue voice (Own) are rated as significantly more natural than those generated with the commercial voice (Ingmar). There is no significant difference in the clearness of the disfluencies in the two voices. There is no significant difference in identification rates for pauses and hesitations between the two voices. This is no surprise since they both used the pitch and durations from the human recordings and since both transitions into silence and fillers sounds are present in both voices. The voice of Own was slightly better at conveying truncations, which could be explained by the fact that truncated diphones were added into this voice, and that these were not present in the voice of Ingmar. The test subjects also listened to a resynthesis of disfluent utterances that were either controlled by a recording from the real interaction or from the reading of a dialogue transcription. For Ingmar the subjects were able to significantly detect the acted version when asked which of the two the acted version was. For some reason this result was not significant for the voice of Own.

It was hard to get significant results for the detection of played disfluencies, especially in the cases where natural and acted disfluencies had been moved into read utterances. The real_to_read_sentences and the played_to_read_sentences are a mixture of the real_sentences and the played_sentences, thus giving no clear distinction between the different sounds and therefore lead to the confusion of the listener. A better approach could have been to stick to the two main recording methods.

Another reason for these ambiguous results can depend on the formulation of the questions. The aim of the listening test was to investigate the clearness and the naturalness of the synthesized disfluency, not the voice quality of the synthesis. There are suspicions that the listeners assessed the quality of the sound in total though the questions were of the kind;

How natural does the disfluent disfluency sound?

Further, the goal was to investigate which method and also which voice gave more natural and spontaneous sounding speech. When asked which disfluency sounded more natural, evidently the listener judged the disfluency through the voice. On the other hand, if the question had been formulated differently, like for example;

How spontaneous does the disfluency sound?

Then the results may have been different. When creating the test it was expected that the listener will may concentrate on the voice as an entity and not just on the parts that are of interested for the project, since the naive user has an untrained ear for this kind of distinctions. Either way, it was decided to run with the original setup of test but at the same time keeping in mind that the answers can contain some bias.

The information given to subjects may have been poor and insufficient in guiding them in answering the questions as planned. A definition of the different types of disfluencies should maybe have been handed out and the different parts of the test should have been more thoroughly explained. Some subjects thought that when hearing the same sound file twice it
meant that the disfluency was repeated, thus classified as repetition. They had problems accepting that every sentence or sound is supposed to be judged independently of the others. Moreover, they thought that the test was too long and tiresome. They got tired by listening to the same sentences over and over again. The listener had been given the opportunity to listen to the files as many times as they wanted. That was a mistake since some could repeat the same sentence up to seven times and that probably confuses more than clarifies anything for the listener. A limit of maximum three times should have been implemented because if the listener does not hear anything the first three times then he won’t hear anything even though he listens to it hundred times.

The upside of the project is that it gives a head-start to how to approach the questions dealt with in this study. In future work more time should be devoted to the dialogue recording and Ingmar. Thou sounding very mechanical the voice is very intelligible, thus if some modifications are made to the sound quality it will probably become the best alternative for synthetic voices.
6 Conclusion

The purpose of this master thesis is to configure a recording method that give raise to naturally provoked disfluent speech. The idea behind this is to implement synthetic voices that can imitate the features of the human’s spontaneous speech. In order to reach the goals of this work, three hypotheses were investigated:

1. Synthesis generated for use in spoken dialogue systems sounds better if they are based on speech corpora provoked in a dialogue situation rather than in laboratory read.
2. Synthesis of disfluent speech can be generated by using two methods; speech which is recorded in a dialogue or by played disfluencies (laboratory read).
3. Disfluencies can be elicited by putting up a role-play that is performed by two individuals, whom are restricted to a specific domain. The elicitation is performed by a third person that eavesdrops on the conversation that takes place between the two role-players and sends them instructions.

The experiment was completed in two parts: A data collection and a listening test. In the data collection two different recording methods were used. The first was based on a human-human dialogue and tested all three hypotheses. The second method was a based on reading a pre-written manuscript and thus tested only the first and the second hypothesis. In the second part of the project a listening test was performed by 16 subjects in order to verify whether or not the hypotheses are true.

The evaluations of the recording sessions showed that the most preferred recording method, by the interlocutors, was the dialogue setup. They had more fun recording the first and second recording sessions due to the reasons that:

1. They received feed-back from a conversational partner
2. Good acting skills were not required
3. They had free reins in choosing their words and how they wish to express them-selves,
4. It was fun and had a somewhat peculiar scene set-up
5. A good theme for the discussion was chosen

The second method, which was based on reading dialogue sentences, was considered a bit boring due to the fact that one interlocutor had to read about 60 sentences. In this scenario the reader was constrained to the written dialogue manuscript and no feed-back from a conversational partner. The recording was plain and straight forward and required good acting skills in order to give the impression of naturally provoked disfluencies. When speaking in a dialogue the disfluent speech was provoked spontaneously primarily, because the speaker gained feed-back from the opponent. Secondly, because the interlocutors received dynamic instructions from the wizard as the conversation went along. Many instructions surprised the participants and that contributed with spontaneous disfluencies. The dialogue session followed the same structure as would a regular discussion between a shopkeeper and customer do. The reasons for why the interlocutors preferred the first method rather than the second method could have been that they felt that they could relate to the situation and that they found it to be very stereotypical for a sell-and-buy interaction.

When the data collection was completed and the listening test was performed the results were gathered and evaluated;
Hypothesis 1:
Disfluencies generated using data from a dialogue are rated as significantly more natural using Own than those generated with the commercial voice (Ingmar). There is no significant difference in the clearness of the disfluencies between the two voices.

Hypothesis 2:
The test subjects listened to a resynthesis of disfluent utterances that were either controlled by a recording from the real interaction or from the reading of a dialogue transcription. For Ingmar the subjects were able to significantly detect the acted version when asked which of the two the acted version was. For some reason this result was not significant for the voice of Own. Truncations generated in Own are rated significantly more natural than the ones synthesized with Ingmar (Kruskal Wallis p=0.001). This can be explained by the fact that truncated diphones were added into the voice of Own, and that these were not present in the voice of Ingmar.

Hypothesis 3:
The disfluency rate while using the instructions was higher than in the previous DEAL conversations. Unfortunately these result are not significant (chi square p=0.2) when only studying two recordings.

In some occasions, listeners identified disfluencies that were actually not there. One example is that though no repairs were included in the listening test many subjects implied that they heard one. While studying figure 4.8, in section 4.2.4, attention should be drawn to the number of identified respectively occurred pauses. The number of occurred pauses was higher than the number identified. That is, pauses were at several occasions confused with other disfluency types as were repetitions. Although a listener could easily hear when an interruption was introduced, the identification of the type of disfluency was not as easy.
7 Future Research

The field of spontaneous speech and dialogue systems is very broad and as mentioned in many parts in the report many researchers before me have done a very good job in doing good findings in this area of research and they’ve also been very generous to sharing their results with us. The results from this project are a bit vague but they can work as a head-start for further eager researchers who would like to dig deeper in this field of work and hopefully take these findings to another level. This project was small for its’ content and the results of the project are based on a small number of contesters. In the future one can choose only one method to run and maybe build a prototype for testing on humans. It’s advisable if the user-test is more quantitative.
8 Acknowledgement

I would like to start by thanking my supervisor, assistant professor Joakim Gustafson at the Department of Speech, Music and Hearing at KTH, for his support and guidance in this project. Big thanks go also to all interlocutors that took part of the experiment, both those that participated in the recordings and those that performed the listening test. I would also like to thank the department of Speech Communication at KTH for giving me chance of being a part of such engaging and fascinating work which I hope will open new doors for the technology of spoken dialogue systems. Many of the colleges at the department have showed a lot of interest and support all through the project, so thank you all for that. Last but not least I’d would like to thank each and every member of my family for their support and endurance. A special thank goes to my father in heaven for his guidance and love.
9 References

9.1 Articles and Books


Corley, M., and Hartsuiker, R. J., *Hesitation in speech can...um...help a listener understand*, School of Philosophy, Psychology and Language Sciences and Human Communication Research Centre at the University of Edinburgh and Department of Experimental Psychology at the Ghent University.


Hjalmarsson A (2002), Evaluating AdApt, a multi-modal conversational, dialogue system using PARADISE; nov. 2002 in Stockholm; Master’s Thesis in Speech Technology, Department of Speech Music and Hearing, KTH Royal Institute of Technology.


Phonetics of talk-in-interaction: a study of the interactional functioning of phonetic detail in everyday talk.


Shriberg, E., Phonetic Consequences of Speech Disfluency, Speech Technology and Research Laboratory, SRI International, Menlo Park, CA.


9.2 Homepages

Black, A. W. (2002), *Unit Selection Synthesis*,


Edlund, J. and Beskow, J., *Pushy vs meek – using avatars to influence turn-taking behavior*,

Hjalmarsson, A., *On cue — additive effects of turn-regulating phenomena in dialogue*,


Jönsson, A. *Dialogsystem*, 2007-01-02, http://sprakteknologi.se/vad-aer-


Mcfadyen, H., *What is an Interjection?*,


Taltekologikursen HT 2007, *Text-till-talsyntes (TTS2)*


9.3 Tables

IPA (The International Phonetic Association), *The International Phonetic Alphabet* (revised to 2005), [http://www.langsci.ucl.ac.uk/ipa/ipachart.html](http://www.langsci.ucl.ac.uk/ipa/ipachart.html), © 2005 IPA.


9.4 Dialogue Systems

Swedbank telefonbank, 0771 22 11 22.

Telia, telnr 90200
[https://www.tewss.telia.se/privat/kundservice/kontaktatelia/viatelefon/?sl=teliase_kservice_kontakta_tele](https://www.tewss.telia.se/privat/kundservice/kontaktatelia/viatelefon/?sl=teliase_kservice_kontakta_tele).

9.5 Further reading

Carlson, R., Gustafson, K. and Strangert, E. *Cues for Hesitation in Speech Synthesis*, A cooperation between CSC, Department of Speech, Music and Hearing, KTH in Stockholm; Acapela Group Sweden AB, Solna and the Department of Comparative Literature and Scandinvian Languages, Umeå University (Sweden).


Gravano, A (2009), *Turn-Taking and Affirmative Cue Words in Task-Oriented Dialogue*, COLUMBIA UNIVERSITY, Submitted in partial fulfillment of the requirements for the degree of Doctor of Philosophy in the Graduate School of Arts and Sciences.


Lickely, R. J. *Juncture cues to disfluency*, Department of Linguistics and Human Communication Research Centre, University of Edinburgh, UK.


10 Wordlist

*ASR* stands for Automatic Speech Recognition.

*Cue phrases* are expressions such as “I mean” and “well”.

*Discourse marker* is a phrase or expression that has no grammatical or syntactic function and it does not intervene with the context of the sentence.

*Interjections* are parts of speech that grammatically have no real connection with the rest of the sentence, for example filled pauses. It expresses emotions and a state of mind.

*Intonational phrase* are phrases or expressions that are stressed.

*Verbal speech* is speech which follows the grammatical rules of a language.

*Non-verbal* speech is expressions that don’t follow any specific grammatical or prosodic set-up.

*Phrase* is a group of words that function as a whole.

*Prolonged* are drawn out durations.

*Repair* is a correction in speech.

*Retrace* is the word which is repeated in speech repair and is located before the alteration.

*Revision* stands for modification of what one has said.

*SDS* is an abbreviation for Spoken Dialogue Systems.

*Speech disfluency* is an interruption in the flow of speech.

*Truncated word* is a word that has been cutoff.

*Utterance* is a complete unit of speech in spoken language.
### Scenario design

<table>
<thead>
<tr>
<th>Instruktioner</th>
<th>Mål/syfte</th>
<th>Kund</th>
<th>Försäljare</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>1</strong></td>
<td></td>
<td></td>
<td>Ser en bild på en leksaks affär</td>
</tr>
<tr>
<td><strong>2</strong></td>
<td>Ge en känsla av att befinna sig i en leksaksaffär</td>
<td>ser en dörr öppnas och kunden går in i ett rum fyllt med leksaker</td>
<td>Hör en dörr öppnas Fotsteg kommer allt närmare</td>
</tr>
<tr>
<td><strong>3</strong></td>
<td>Sätta igång samtalet</td>
<td></td>
<td>Starta samtalet som en vanlig försäljare</td>
</tr>
<tr>
<td><strong>4</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>5</strong></td>
<td>Inrikta sig på ett tema</td>
<td>Ge dina önskemål</td>
<td></td>
</tr>
<tr>
<td><strong>6</strong></td>
<td>Göra F. Lite osäker</td>
<td>Visa missnöje</td>
<td>Placera ut 4 alt.</td>
</tr>
<tr>
<td><strong>7</strong></td>
<td>Framkalla tvekan och paus hos K</td>
<td>Bilderna försvinner väldigt snabbt</td>
<td>Placera 4 nya alt</td>
</tr>
<tr>
<td><strong>8</strong></td>
<td>K och F ser olika många produkter</td>
<td>Ser endast 3 av fyra</td>
<td></td>
</tr>
<tr>
<td><strong>9</strong></td>
<td></td>
<td>Visa intresse för nr 2 från höger</td>
<td></td>
</tr>
<tr>
<td><strong>10</strong></td>
<td>Framkalla tvekan hos F</td>
<td>Bete dig som om du har vunnit på lotto</td>
<td></td>
</tr>
<tr>
<td><strong>11</strong></td>
<td>F upptäcker att ett misstag har skett</td>
<td>Beskriv produkten</td>
<td></td>
</tr>
<tr>
<td><strong>12</strong></td>
<td>Framkalla stakningar hos K</td>
<td>Bete dig så som om du har förlorat det bästa i livet</td>
<td></td>
</tr>
<tr>
<td><strong>13</strong></td>
<td>Föra vidare diskussionen Framkalla rättningar och kanske upprepningar hos F</td>
<td></td>
<td>Övertyla kunden att köpa något annat istället-prata snabbt</td>
</tr>
<tr>
<td><strong>14</strong></td>
<td>Framkalla avbrott hos F</td>
<td>Var en jobbig kund</td>
<td></td>
</tr>
<tr>
<td><strong>15</strong></td>
<td>Göra K nervös så att han tvekar, pausar, upprepår</td>
<td>Bete dig ohyfsat- ha humörsvävningar</td>
<td></td>
</tr>
<tr>
<td><strong>16</strong></td>
<td>Framkalla avbrott hos bådeK och F</td>
<td>Prata på utan paus om hur du vill att produkten du köper ska vara-envisa</td>
<td>Avbryt kunden hela tiden genom att berätta hur bra de andra produkterna är</td>
</tr>
<tr>
<td><strong>17</strong></td>
<td>Framkalla tvekan hos K</td>
<td>Bete dig osäkt</td>
<td></td>
</tr>
<tr>
<td><strong>18</strong></td>
<td>Göra F nervös-framkalla rättningar, tvekan, upprepningar</td>
<td>Ring din fru varje gång försäljaren kommer med ett nytt förslag-prata lugnt och fint utan stress</td>
<td>Kom med ett nytt förslag snabbt så fort K vill ringa sin fru</td>
</tr>
<tr>
<td><strong>19</strong></td>
<td>Först och främst pauser hos K</td>
<td>Var gnällig och prata om hur orättvist livet är-prata långsamt</td>
<td></td>
</tr>
<tr>
<td><strong>20</strong></td>
<td>Rättningar hos K Kanske avbrott hos F</td>
<td>Andra åsikt hela tiden. Prata</td>
<td></td>
</tr>
<tr>
<td>23</td>
<td>Tvekan och stakningar hos K</td>
<td>Välj ut en produkt</td>
<td>Var missbelåten med valet</td>
</tr>
<tr>
<td>24</td>
<td>Få K att pausa samt tveka</td>
<td>Fråga efter priset på en del produkter som du blir intresserad av</td>
<td>Ge helt sanslöst och orimligt höga priser på alla produkter K väljer</td>
</tr>
<tr>
<td>25</td>
<td>Turn-taking och kanske en del stakningar</td>
<td>Var snål</td>
<td>Var girig och bestämd</td>
</tr>
<tr>
<td>26</td>
<td>Turn-taking, Upprepningar hos F, Pauser hos K</td>
<td>Hota med att inte köpa något varje gång försäljaren inte går ned till ett pris som du tycker låter rimligt</td>
<td>Håna kunden varje gång han hotar med att inte köpa något</td>
</tr>
<tr>
<td>27</td>
<td>Gå vidare i diskussionen</td>
<td></td>
<td>Var lite mer tillmötesgående</td>
</tr>
<tr>
<td>29</td>
<td>Tvekningar och upprepningar hos K, Pauser hos F</td>
<td>Du velar-har belutsdilemma</td>
<td></td>
</tr>
<tr>
<td>30</td>
<td>Avbrott hos K, Upprepningar hos F</td>
<td>Prata som en psykolog</td>
<td>Bete dig helt hysteriskt</td>
</tr>
<tr>
<td>31</td>
<td>Avsluta samtalen</td>
<td>Du smitter iväg</td>
<td></td>
</tr>
</tbody>
</table>
11.2 Appendix B

Persona description

Kund

Daniel Hagelström är en man i sina 40. Han är gift och har 2 barn, David och Maria. David är 10 år och Maria fyller 5 år om två dagar. Daniel och Birgitta har varit gifta i tolv år, men de har varit tillsammans i 20 år. Familjen Hagelström bor i en villa i Salem kommun. De har bott där i 10 år och stortrivs.


En av Daniels svagpunkter är att han har problem med fattta beslut och hålla sig till dem. Han är väldigt beroende av sin fru och kan inte besluta något innan han har pratat med henne. Speciellt på senare tid har han visat tydliga tecken på dåligt självförtroende. Att inte ha ett jobb att gå till varje morgon har verkligen satt sina spår i honom. Daniel känner sig vårdlös och som en börda för sin familj. Han går igenom en stark depressionsperiod och ingen vet om han kommer att lyckas att ta sig igenom den.

Normalt är Daniel en glad och optimistisk person som älskar livet. Han brukar alltid hitta på nya äventyrliga och kreativa utflykter för familjen att göra tillsammans. Men sedan han fick nyheten om att han inte kan fortsätta att jobba på företaget har han varit väldigt inständig, gnällig och pessimistisk. Han tycker att allt är tråkigt och vårdlös. Han har även blivit lite mer varsam med hur och på vad han spenderar sina pengar.

Försäljare


Hasse har en flickvän, Valentina. Hasse träffade Valentina på en kryssning till Åbo för 2 år sedan. För ett halvår sedan bad Hasse Valentina att flytta in till honom eftersom han tyckte att
det var väldigt jobbigt med ett långdistansförhållande. Valentina bodde i Skåne innan hon flyttade in med Hasse. De bor i en trea i kommundelen Rönninge.

Hasse är 1.80 m lång, har lagom kraftig kroppbyggnad, har ljusbrunt snaggat hår och mörkgrön ögonfärg. Han har ett fyrkantigt ansikte med fina manliga drag. Han är medveten om sitt charmiga utseende, vilket utstrålar självförtroende. Hans självförtroende, kombinerat med hans finslipade sociala kompetens och övertalningsförmåga, har lett till att folk oftast gör som han säger. Detta har hjälpt honom mycket i sina affärer. Mot kvinnliga köpare utnyttjar han sin charm för att övertala dem att köpa hans produkter, mot män, däremot, är han mer bestämd och sakkunnig.

I affärer, som nämnt tidigare, är Hasse väldigt hård och det är med illvilja han går ned i pris för att sälja en vara, men det händer ibland att han ger sig och låter kunden vinna från tid till tid. Mårker han att kunden envisar mycket eller att dem är upprätta när dem säger att de inte har råd kan han tänka sig att kompromissa, men inte utan strid.
11.3 Appendix C

**Steady instructions**

**Kund**

- Daniel är en person som tycker om att höra andras åsikter, speciellt försäljaren han handlar av, innan han bestämmer sig för att köpa något. Han brukar ställa alla möjliga frågor om produkten och dess egenskaper. David är även en stor beundrare av varförföljdfrågor, t.ex varför tycker du att den gröna färgen är finare än den röda eller varför är A den första bokstaven i alfabetet osv.
- Daniel har 300 kr till sin disposition och alla prisförslag han får är exklusive moms, vilket innebär att han måste lägga till 25% på priset för att få ut den totala priset för varan.

**Försäljare**

- Företaget som Hasses jobbar för och är delägare i har en vinstpolicy samt en förhandlingsstrategi. Det första prisförslaget som ges till kunden ska bestå av 75 % vinst och det sista ska bestå av minst 25 % vinst. Under förhandlingsskedet får försäljaren endast minska priset med 10 % åt gången, vilket innebär att vinsten minskar med ungefär 3 % åt gången.
11.4 Appendix D

Manus 1

F: Inled samtalet som en vanlig försäljare

K: Ge dina önskemål

Visar upp scenario 1 – Likadana produkter

F: Presenterar de olika produkterna lite kort

K: Missnöjd med alla

Visar upp scenario 4 – likadana produkter men blandade (trasiga, hela, bosch)

K: Välj en produkt

(Tar bort den som köparen vill ha)

F: Varan är slutsåld

Meddela köparen att produkten som han valde ut såldes precis till ngn annan

K: Var besviken och dyster

F: Övertala köpare att köpa en av de andra produkterna

K: Säg nej till alla förslag

Scenario 6

F: Sälj exempelvis blå bil/gul docka/ lila nalle/rosa docka

Väntar på reaktion från köparen.

Ändra bilderna

Scenario 5

K: Var gnållig

F: Uppför dig ohyfsat – ha humörsvängningar

K: Ställ många frågor

F: Ställ många motfrågor

F: Sälj rosa bil

K: Bete dig osäkert

K: Du velar – har beslutsdilemma

F: Var otålig och irriterad

F: Sälj ngt (välj ut ngt)

(Prata långsamt)

K: Säg inget vid paus

69
F: Sälj monsterägget
K: Säg va hela tiden
F: Sälj blå nalle
K: Prata på utan paus
F: Avbryt hela tiden
K: Köp röd docka
F: Mumla

Scenario 2
F: Sälj ngt
K: Be om ursäkt hela tiden
F: Tala om för köparen vad de olika produkterna kostar utan att han frågar om det (Ge orimligt höga prisförslag)
K: Var gnällig samt prata långsamt om hur orättvist livet är
F: Upprepa allt köparen säger

(Välj ut ngr produkter samt prissätt dem)
K: Hota med att inte köpa ngt varje gång försäljaren vägrar gå ner i pris
F: Var girig och bestämd
F: Övertala kunden snabbt att köpa en av Bosch produktarna
K: Köp en rosa docka
F: Sälj fisken med tornet
K: Ställ svåra emotionella frågor
K: Köp fisken med tornet
F: Var missnöjd med köparens val av produkt
F: Sälj den enögda rosa dockan
K: Var kräsen
F: Prata snabbt och osammanhängande
F: Sälj lila docka
K: Köp lila docka
K: Ge ett prisförslag som är högre än försäljarens – insistera
F: Pruta ner priset
F: Börja avsluta affären
K: Börja avsluta affären
K: Du ångrar ditt beslut och försöker att slingra dig undan köpet
F: Bli hysterisk
K: Lägg på
11.5 Appendix E

Manus 2

F: Inled samtalet som en vanlig försäljare

K: Ge dina önskemål

Visar upp scenario 1 – Likadana produkter

F: Presenterar de olika produktarna lite kort

K: Du är inte riktigt nöjd med förslagen som försäljaren kommer med

Visar upp scenario 4 – likadana produkter men blandade (trasiga, hela, bosch)

F: Övertala köparen att köpa någon av dina produkter

K: Säg nej till alla förslag

Scenario 6 – olika produkter, 4 st

F: Sälj exempelvis blå bil/gul docka/ lila nalle/rosa docka

Väntar på reaktion från köparen.

Ändrar bilderna manuellt när kund reagerar

K: Var gnällig – prata om hur orättvist livet är

F: Upptör dig ohyfsat – ha humörsvängningar

K: Ställ många frågor om produkterna du ser på skärmnen

F: Ställ många motfrågor

Scenario 5 – enbart bilar, 5 st

F: Sälj rosa bil

K: Du har beslutsdilemma

F: Var riktigt otåligen och irriterad

Scenario 7 – monsterägget och en trasig blå nalle

F: Sälj monsterägget

Prata långsamt

K: Lätsas att du inte hör vad säljaren säger

F : Sälj blå nalle

K: Köp röd docka

Scenario 2 – likadana men enbart dockor eller bosch produkter (trasiga, hela och bosch)

F: Sälj vilket som förutom den rosa dockan

K: Du fastnar för den rosa dockan
F: Tala om för köparen vad de olika produktarna kostar utan att han frågar om det – ändra till högre priser så fort säljaren lovat köparen ett pris – ändra sedan till låga priser igen

K: Köp den trasiga gröna bilen
F: Du vägrar sälja den trasiga gröna bilen

Scenario 3 – likadana men blandade produkter (4 st)

F: Sälj fisken med tornet
K: Köp fisken med tornet

F: Var missnöjd med köparens val av produkt – sälj den enögda rosa dockan istället – visa den trasiga rosa dockan
K: Upprepa allt säljaren säger
F: Prata på om hur du träffade din flickvän
K: Få försäljaren att berätta lite om de olika produktarna

Visa den lila dockan
F: Sälj lila docka
K: Köp lila docka

Prissätt dockan med 55kr
K: Ge ett prisförslag som är högre än försäljarens – insistera
F: Pruta ner priset
K: Hota med att inte köpa något om försäljaren inte går upp till det pris du vill
F: Börja avsluta affären
K: Börja avsluta affären

K: Du ångrar ditt beslut och försöker att slingra dig undan köpet genom att förolämpa säljaren
F: Lägg på
K: Lägg på
11.6 Appendix F

Segments chosen for recording session 3

Inspelning 1

1. Vi kan gå omkring lite i butiken och kolla på (paus) tomma hyllor här nu ska vi se
   - vill du gå på (paus) bio med mig

2. ähh du kan till och med få den för hundra kroner
   - ähhh drottningen vill nog ha den röda dräkten

3. den där rosa dockan e (paus) e ju lite sällskapssjuk (paus) ähhh så jag måste nog säga att
   - ja men han e e jätte sjuk ähhh så han stannar nog hemma idag

4. äh de e alltså en leks en leksaksbil (paus) va sa du
   - äh du kan få en leks en leksaksbil (paus) för hundra kronor

5. det här e ju det här e ju en bil till dockan
   - det här e ju det här e ju ett citat taget från hamlet

6. ja men va va sk (avbrott) nä men du kan ju (avbrott) varför ska jag ge dig rabatt
   - hur kan sanningen va va sk (avbrott) när hörde du (avbrott) va menar du

7. vi kanske kan göra så här istället ähhh den lila bilen som du ser så har även den punka på på
   vänster bak
   - det är förbjudet att ähhh duscha på på vardagar

8. alltså det kostar femton miljoner de e liksom mm de e inte så mycket å göra åt
   - jag menar de e liksom mm de e inte så svårt som du tror

9. ja de e inte som ja alltså bb jag måste ju bb de liksom hh inte gå med vinst men jag kan inte gå
   med fö g med förlust åhh men ja menar
   - som egenföretagare så bb jag vet att det är svårt att hålla reda på alla men du bb litar lite för
     mycket på dina anställda sam hh insåg du inte att ditt företag började gå med fö g med förlust
     åhh mycket pengar har försvunnit utan någon riktig förklaring

10. ja mena (paus) ja vet inte varför den framkallar obehag hos dig
    - var finns böckerna som jag köpte till anna (paus) jag hade ställt dem på bordet

11. a alltså ja förstår inte jag förstår inte
    - varför säger du så ja förstår inte jag förstår inte
    - hurdå ja förstår inte jag förstår inte

Inspelning 2

1. Hej ähhb ja öh antar jag ska sälja lite grejer till dig
   - vilken kaj ähhb ja öh använder du till din yacht

2. ähhhh men vi kan väl säga så här va att h ja tänkte faktiskt bör aasj ja fø va bra för jag hade
   tänkt
-ähhhh man måste vara väldigt h jo försiktig när man ska aasj ja fö visa barnen hur man ska gå tillväga när man går över gatan

3. ähhh å den e bb (avbrott - paus) väldigt inne på bi (avbrott - paus) öhh de e de e lite
-ähh åsnan är bb (avbrott - paus) vacker men den är li (avbrott - paus) öhh de e de e lite för lat

4. nääh de tycker ja inte den e e den e faktiskt begagnad
-ääh david har flyttat till en ny lägenhet belägen i skåne den ä den e fruktansvärt gammal

5. jo men a ja kan bara säga att om du kö (avbrott) istället för den rosa dockan så kan du köpa (paus) inte mindre än tre stycken nallar (paus ) å dessutom ha pengar kvar på fickan
-joh men ja ja kanske borde kö (avbrott) i stallet finns det en oljelampa (paus) inlåst i en träläda för att öppna lådan behöver du skiftnycklar (paus) dom hittar du på en av hyllorna i stugan

6. så de där fattar du de e ändå de e ändå en ganska dyrt det blir ett samlarobjekt me nån form av bb den har ju ett investeringssvärde kan man säga så
-sju de e ändå de e ändå en bra start och två av bb dessa går till välgörenhet

7. ja men de e fortfarande så att v (avbrott)
-jag vet inte hur jag ska förklara detta för dig men arv (avbrott)

8. vet du va ja då tycker att du gör du köper du köper den rosa dockan
-det nybakade brödet du gör du köper du köper doftar jätte gott

9. den går ju på tre g (avbrott) nä den går på tre g den e inte alls lika bra
-den går ju på tre g (avbrott) nä den går på tre g den behöver tre stycken AAA batterier

10. ja nn alltså nääh ja kan inte sälja den fö för nätting överhuvudtaget för att den där dockan längst ner e så (avbrott)
-ja nn alla är säkra inte lika smarta å nääh jag kan inte ljuga för dig du är otroligt dum och blå(avbrott)

11. ännu faktiskt (paus) för nu e du du nu e du konstig faktiskt
-du måste tro på jesus nu när du e döpt (paus) för nu e du du nu nu e du kristen

12. ja tycker at (avbrott) nähå (paus) nä ja tycker du ska köpa dockan faktiskt så så så (avbrott) nä men v b o a
-hoppa på at (avbrott) nähå (paus) nä jag tar tillbaka det jag kan nog inte köra dig också (avbrott) nä men v b o ä

13. jo men (paus) ja e inte oärlig när jag b berättar om den här dockan den e ju faktiskt huf hur bra som helst (paus) å så har du den här dockan h (avbrott)
-varken (paus) juice eller mjölk jag vill inte dricka något alls släng b bort allt huf har ingen lust (paus) att åta eller dricka något alls eller förresten ge mig ett glas wh (avbrott)
11.7 Appendix G

The segments used for recording session 3

Inspelning 1 (Track2)

1. den där rosa dockan e (paus) e ju lite sällskapssjuk (paus)ähhh så jag måste nog säga att
   -ja men han e e jätte sjuk ähhh så han stannar nog hemma idag

2. äh de e alltså en leks en leksaksbil (paus) va sa du
   -äh du kan få en leks en leksaksbil (paus) för hundra kronor

3. ja men va va sk (avbrott) nä men du kan ju (avbrott) varför ska jag ge dig rabatt
   -hur kan sanningen va va sk (avbrott) när hörde du (avbrott) va menar du

4. ja mena (paus) ja vet inte varför den framkallar obehag hos dig
   -var finns böckerna som jag köpte till anna (paus) jag hade ställt dem på bordet

5. a alltså ja förstår inte jag förstår inte
   -varför säger du så ja förstår inte jag förstår inte
   -hurda ja förstår inte jag förstår inte

Inspelning 2 (Track1)

6. nääh de tycker ja inte den e e den e faktiskt begagnad
   -ääh david har flyttat till en ny lägenhet belägen i skåne den är den e fruktansvärt gammal

7. jo men a ja kan bara säga att om du kö (avbrott) istället för den rosa dockan så kan du köpa (paus) inte mindre än tre stycken nallar (paus) å dessutom ha pengar kvar på flickan
   -joh men ja ja kanske borde kö (avbrott) i stället finns det en oljelampa (paus) inlåst i en trälåda för att öppna lådan behöver du skiftnycklar (paus) dom hittar du på en av hyllorna i stugan

8. vet du va ja då tycker att du gör du köper du köper den rosa dockan
   -det nybakade brödet du gör du köper du köper doftar jättegott

9. den går ju på tre g (avbrott) nä den går på tre g den e inte alls lika bra
   -den går ju på tre g (avbrott) nä den går på tre g den behöver tre stycken AAA batterier

10. ja nn alltså nääh ja kan inte sälja den fö för nånting överhuvudtaget för att den där dockan längst ner e så (avbrott)
    -ja nn alla är såklart inte lika smarta å nääh jag kan inte ljuga för dig du är otroligt dum och blå (avbrott)

11. ännu faktiskt (paus) för nu e du du nu e du konstig faktiskt
    -du måste tro på jesus nu när du döpt (paus) för nu e du du du nu e du kristen
12. **jo men (paus)** ja e inte oärlig när jag b berättar om den här dockan den e ju faktiskt huf hur bra som helst (paus) å så har du den här dockan h(avbrott)
-varken (paus) juice eller mjölk jag vill inte dricka något alls släng b bort allt huf har ingen lust (paus) att äta eller dricka något alls eller förresten ge mig ett glas wh (avbrott)
11.8 Appendix H

Manus 1

_Analyzed manuscript after the first recording_

F: Inled samtalet som en vanlig försäljare
K: Ge dina önskemål

_Visar upp scenario 1 – Likadana produkter_

F: Presenterar de olika produktorna lite kort
K: Missnöjd med alla – gick vidare lite för snabbt – lite stakningar kanske (F)

_Visar upp scenario 4 – likadana produkter men blandade (trasiga, hela, bosch)_

K: Välj en produkt
_(Tar bort den som köparen vill ha)_

F: Varan är slutsåld

Meddela köparen att produkten som han valde ut såldes precis till ngn annan – funkade inte riktigt

K: Var besviken och dyster - använde inte

Bra F: Övertala köpare att köpa en av de andra produktorna – paus, avbrott, tvekan (F), upprepning,

Bra K: Säg nej till alla förslag – tvekan, paus (K)

Scenario 6

F: Sälj exempelvis blå bil/gul docka/ lila nalle/rosa docka

_Väntar på reaktion från köparen. Framkallde tvekan och paus hos F istället för K_

_Antingen ändrar vi bilderna manuellt eller så väljer vi en ny scen_

Scenario 5

K: Var gnällig

Bra F: Uppför dig ohyfsat – ha humörsvängningar – tvekan, paus, upprepning (K), avbrott, tvekan (F)

Bra K: Ställ många frågor - tvekan, paus (F), tvekan, upprepningar (K)

Bra F: Ställ många motfrågor

F: Sälj rosa bil

Bra K: Bete dig osäkert – tvekan, paus (K), tvekan, paus, upprepningar (F)

Bra K: _Du velar – har beslutsdilemma_ – upprepning (K), upprepning, tvekan, rättning (F)
Bra F: **Vår otålig och irriterad**

F: Sälj ngt (**välj ut ngt**)  
Prata långsamt – tvekan, paus, upprepning, falsk start (F)  
K: Säg inget vid paus – upprepning (K)  
F: **Sälj monsterägget** – paus, tvekan, upprepning, rättning (F)  
K: **Säg va hela tiden**  
F: Sälj blå nalle  
K: **Prata på utan paus** – tvekan, paus, avbrott (K), avbrott (F) – pratade i mun på varandra  
F: **Avbryt hela tiden**  
K: Köp röd docka  
F: Mumla – funkade inte riktigt – tvekan, upprepning, paus (F)

**Scenario 2**

F: **Sälj ngt** – avbrott, tvekan, paus, upprepning (F & K)  
K: **Be om ursäkt hela tiden**  
F: Tala om för köparen vad de olika produkterna kostar utan att han frågar om det (**Ge orimligt höga prisförslag**) – tvekan, upprepning, avbrott, rättning (F), paus, upprepning, tvekan (K)  
K: **Var gnällig samt prata långsamt om hur orättvist livet är** – paus, tvekan (K)  
F: **Upprepa allt köparen säger** – avbrott, tvekan, paus (F)  

(Välj ut ngr produkter samt prissätt dem)  
K: Hota med att inte köpa ngt varje gång försäljaren vägrar gå ner i pris – uprep., tvekan (K), tvekan, paus, falska starter (F)  
F: Var girig och bestämd  
F: Övertala kunden snabbt att köpa en av Bosch produkterna  
K: **Köp en rosa docka** – avbrott (K), avbrott, tvekan, paus, upprepning (F)  
F: **Sälj fisken med tornet**  
K: Ställ svåra emotionella frågor – tvekan, upprepn. (K), tvekan, paus, avbrott (F)  
K: Köp fisken med tornet  
F: Var missnöjd med köparens val av produkt – upprep., paus, tvekan, avbrott (F), avbrott (K)  
F: Sälj den enögda rosa dockan

79
K: Var kräsen – avbrott (K), tvekan, upprepning, avbrott (F)

F: Prata snabbt och osammanhängande – funkade ej

F: Sälj lila docka

K: Köp lila docka

K: Ge ett prisförslag som är högre än försäljarens – insistera – tvekan, paus (F & K)

F: Pruta ner priset - tvekan, paus (K)

F: Börja avsluta affären

K: Börja avsluta affären

K: Du ångrar ditt beslut och försöker att slingra dig undan köpet – tvekan, paus, rättning (K)

F: Bli hysterisk – upprepning, tvekan, rättning (F), avbrott (K)

K: Lägg på
11.9 Appendix I

The test results of part two in the listening test

a) The basic data for figure 4.1a

<table>
<thead>
<tr>
<th>Voice</th>
<th>MyOtyd</th>
<th>Otyd</th>
<th>GaOtyd</th>
<th>GaTyd</th>
<th>Tyd</th>
<th>MyTyd</th>
<th>Grand Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>AMR</td>
<td>2</td>
<td>7</td>
<td>6</td>
<td>20</td>
<td>26</td>
<td>19</td>
<td>80</td>
</tr>
<tr>
<td>Own</td>
<td>3</td>
<td>12</td>
<td>10</td>
<td>33</td>
<td>52</td>
<td>29</td>
<td>139</td>
</tr>
<tr>
<td>Ingmar</td>
<td>2</td>
<td>16</td>
<td>9</td>
<td>33</td>
<td>46</td>
<td>41</td>
<td>147</td>
</tr>
<tr>
<td>Grand total</td>
<td>7</td>
<td>35</td>
<td>25</td>
<td>86</td>
<td>124</td>
<td>89</td>
<td>366</td>
</tr>
</tbody>
</table>

b) The basic data for figure 4.1b

<table>
<thead>
<tr>
<th>Voice</th>
<th>MyOnat</th>
<th>Onat</th>
<th>GaOnat</th>
<th>GaNat</th>
<th>Nat</th>
<th>MyNat</th>
<th>Grand Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>AMR</td>
<td>4</td>
<td>8</td>
<td>23</td>
<td>24</td>
<td>25</td>
<td>84</td>
<td></td>
</tr>
<tr>
<td>Own</td>
<td>3</td>
<td>23</td>
<td>24</td>
<td>42</td>
<td>40</td>
<td>18</td>
<td>150</td>
</tr>
<tr>
<td>Ingmar</td>
<td>22</td>
<td>29</td>
<td>30</td>
<td>37</td>
<td>24</td>
<td>14</td>
<td>156</td>
</tr>
<tr>
<td>Grand total</td>
<td>25</td>
<td>56</td>
<td>62</td>
<td>102</td>
<td>88</td>
<td>57</td>
<td>390</td>
</tr>
</tbody>
</table>

c) The basic data for figure 4.4

<table>
<thead>
<tr>
<th>Voice</th>
<th>Number of cutoff</th>
<th>Number of repetition</th>
<th>Number of pause</th>
<th>Number of hesitation</th>
<th>Number of repair</th>
</tr>
</thead>
<tbody>
<tr>
<td>AMR</td>
<td>20</td>
<td>27</td>
<td>35</td>
<td>37</td>
<td>7</td>
</tr>
<tr>
<td>Own</td>
<td>53</td>
<td>48</td>
<td>55</td>
<td>70</td>
<td>13</td>
</tr>
<tr>
<td>Ingmar</td>
<td>82</td>
<td>51</td>
<td>52</td>
<td>53</td>
<td>19</td>
</tr>
<tr>
<td>Grand total</td>
<td>155</td>
<td>126</td>
<td>142</td>
<td>160</td>
<td>39</td>
</tr>
</tbody>
</table>
11.10 Appendix J

The test results of part three in the listening test

a) The basic data for figure 4.2

<table>
<thead>
<tr>
<th>Mixtype</th>
<th>Played</th>
<th>Real</th>
<th>Grand total</th>
</tr>
</thead>
<tbody>
<tr>
<td>played_sentence</td>
<td>72</td>
<td>48</td>
<td>120</td>
</tr>
<tr>
<td>played_to_read_sentence</td>
<td>56</td>
<td>64</td>
<td>120</td>
</tr>
<tr>
<td>real_dialogue</td>
<td>48</td>
<td>72</td>
<td>120</td>
</tr>
<tr>
<td>real_to_read_sentence</td>
<td>64</td>
<td>56</td>
<td>120</td>
</tr>
<tr>
<td>Grand total</td>
<td>240</td>
<td>240</td>
<td>480</td>
</tr>
</tbody>
</table>

b) The basic data for figure 4.5

<table>
<thead>
<tr>
<th>Voice</th>
<th>Number of hesitation</th>
<th>Number of pause</th>
<th>Number of cutoff</th>
<th>Number of repetition</th>
<th>Number of repair</th>
</tr>
</thead>
<tbody>
<tr>
<td>Ingmar</td>
<td>122</td>
<td>135</td>
<td>83</td>
<td>80</td>
<td>29</td>
</tr>
<tr>
<td>Own</td>
<td>128</td>
<td>127</td>
<td>67</td>
<td>88</td>
<td>22</td>
</tr>
<tr>
<td>Grand total</td>
<td>250</td>
<td>262</td>
<td>150</td>
<td>168</td>
<td>51</td>
</tr>
</tbody>
</table>
11.11 Appendix K

a) Basic data for figure 4.3; part four in the listening test

<table>
<thead>
<tr>
<th>Voice</th>
<th>Break</th>
<th>Other</th>
<th>Same</th>
<th>Grand total</th>
</tr>
</thead>
<tbody>
<tr>
<td>Ingmar</td>
<td>break</td>
<td>24</td>
<td>21</td>
<td>45</td>
</tr>
<tr>
<td></td>
<td>pause</td>
<td>18</td>
<td>27</td>
<td>45</td>
</tr>
<tr>
<td>Ingmar total</td>
<td></td>
<td>42</td>
<td>48</td>
<td>90</td>
</tr>
<tr>
<td>Own</td>
<td>break</td>
<td>99</td>
<td>171</td>
<td>270</td>
</tr>
<tr>
<td></td>
<td>pause</td>
<td>72</td>
<td>108</td>
<td>180</td>
</tr>
<tr>
<td>Own total</td>
<td></td>
<td>171</td>
<td>279</td>
<td>450</td>
</tr>
<tr>
<td>Grand total</td>
<td></td>
<td>213</td>
<td>327</td>
<td>540</td>
</tr>
</tbody>
</table>

b) Basic data for figure 4.6; part four in the listening test

<table>
<thead>
<tr>
<th>Break</th>
<th>Type2</th>
<th>Other</th>
<th>Same</th>
<th>Grand total</th>
</tr>
</thead>
<tbody>
<tr>
<td>Break</td>
<td>creaky</td>
<td>13</td>
<td>32</td>
<td>45</td>
</tr>
<tr>
<td></td>
<td>cutoff</td>
<td>53</td>
<td>112</td>
<td>165</td>
</tr>
<tr>
<td></td>
<td>exhalation</td>
<td>18</td>
<td>12</td>
<td>30</td>
</tr>
<tr>
<td></td>
<td>inhalation</td>
<td>28</td>
<td>32</td>
<td>60</td>
</tr>
<tr>
<td></td>
<td>none</td>
<td>11</td>
<td>4</td>
<td>15</td>
</tr>
<tr>
<td>Break total</td>
<td></td>
<td>123</td>
<td>192</td>
<td>315</td>
</tr>
<tr>
<td>Pause</td>
<td>cutoff</td>
<td>27</td>
<td>48</td>
<td>75</td>
</tr>
<tr>
<td></td>
<td>none</td>
<td>63</td>
<td>87</td>
<td>150</td>
</tr>
<tr>
<td>Pause total</td>
<td></td>
<td>90</td>
<td>135</td>
<td>225</td>
</tr>
<tr>
<td>Grand total</td>
<td></td>
<td>213</td>
<td>327</td>
<td>540</td>
</tr>
</tbody>
</table>
### 11.12 Appendix L

*The test results of part five in the listening test*

<table>
<thead>
<tr>
<th>User</th>
<th>id1</th>
<th>best</th>
<th>id2</th>
<th>own_change</th>
<th>id3</th>
<th>ing_change</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>BEST:</td>
<td>own</td>
<td>OWN:</td>
<td>kortare pauser</td>
<td>ING:</td>
<td>låter som en robot</td>
</tr>
<tr>
<td>2</td>
<td>BEST:</td>
<td>ingmar</td>
<td>OWN:</td>
<td>mer angagerad i samtalet</td>
<td>ING:</td>
<td>ordspråkt</td>
</tr>
<tr>
<td>3</td>
<td>BEST:</td>
<td>ingmar</td>
<td>OWN:</td>
<td>gladare, fler ljudnivåer och snabbare</td>
<td>ING:</td>
<td>tydligare</td>
</tr>
<tr>
<td>4</td>
<td>BEST:</td>
<td>Ingmar</td>
<td>OWN:</td>
<td>Han upprepar sig och är otydlig</td>
<td>ING:</td>
<td></td>
</tr>
<tr>
<td>5</td>
<td>BEST:</td>
<td>own</td>
<td>OWN:</td>
<td>tevekan, pausen</td>
<td>ING:</td>
<td></td>
</tr>
<tr>
<td>6</td>
<td>BEST:</td>
<td>own</td>
<td>OWN:</td>
<td></td>
<td>ING:</td>
<td></td>
</tr>
<tr>
<td>7</td>
<td>BEST:</td>
<td>ingmar</td>
<td>OWN:</td>
<td></td>
<td>ING:</td>
<td></td>
</tr>
<tr>
<td>8</td>
<td>BEST:</td>
<td>own</td>
<td>OWN:</td>
<td></td>
<td>ING:</td>
<td></td>
</tr>
<tr>
<td>9</td>
<td>BEST:</td>
<td>own</td>
<td>OWN:</td>
<td></td>
<td>ING:</td>
<td></td>
</tr>
<tr>
<td>10</td>
<td>BEST:</td>
<td>own</td>
<td>OWN:</td>
<td>för mycket tvekande i snacket och det finns ingen sammankoppling på det rösten säger, för många pauser</td>
<td>ING:</td>
<td>för mycket upprepning på vissa ord och tvekande</td>
</tr>
<tr>
<td>11</td>
<td>BEST:</td>
<td>own</td>
<td>OWN:</td>
<td>Ta bort avbrott</td>
<td>ING:</td>
<td>Ta bort avbrott, otydlig, irriterande</td>
</tr>
<tr>
<td>12</td>
<td>BEST:</td>
<td>own</td>
<td>OWN:</td>
<td>lite mjukare</td>
<td>ING:</td>
<td>göra den mer männsklig och snabbheteb ska minskas</td>
</tr>
<tr>
<td>13</td>
<td>BEST:</td>
<td>own</td>
<td>OWN:</td>
<td></td>
<td>ING:</td>
<td></td>
</tr>
<tr>
<td>14</td>
<td>BEST:</td>
<td>own</td>
<td>OWN:</td>
<td></td>
<td>ING:</td>
<td></td>
</tr>
<tr>
<td>15</td>
<td>BEST:</td>
<td>own</td>
<td>OWN:</td>
<td>inte så mycket, men lite bättre röstkvalitet</td>
<td>ING:</td>
<td>inte så mycket faktiskt, men samma här, lite bättre röstkvalitet skulle inte skada + att den ARTIKULERAR lite för tydligt</td>
</tr>
<tr>
<td>16</td>
<td>BEST:</td>
<td>own</td>
<td>OWN:</td>
<td>ändra på tveksamheten</td>
<td>ING:</td>
<td>ta bort det avbrutna som repeeterar sig</td>
</tr>
</tbody>
</table>

id1 stands for which of the voices sounds best  
id2 stands for the changes the subjects wish to do to Own  
id3 stands for the improvements the subjects wish to do to Ingmar