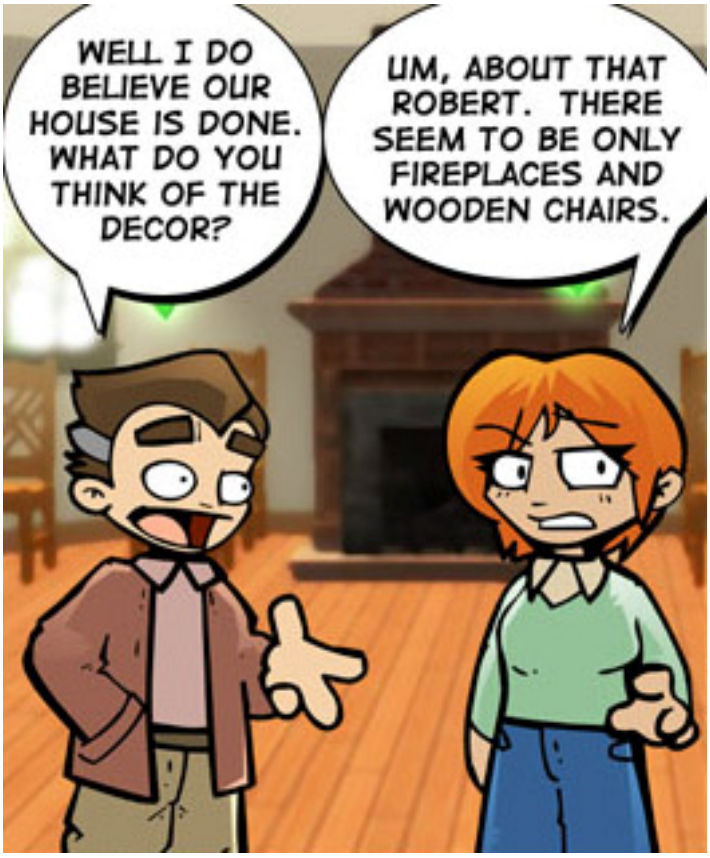
Enhancing Synthetic Speech with Filled Pauses



VG Cats – 128 (www.vgcats.com)

Enhancing Synthetic Speech with Filled Pauses

Submitted in partial fulfillment
of requirements to graduate from
the Department of Computer Science
chair of Human Media Interaction at
the University of Twente on
September 27, 2007

by

Arien S. Kock

Examination Committee:

1st committee member: Zsofia Ruttkay
(zsofi@ewi.utwente.nl)

Arjan van Hessen
(A.J.vanHessen@ewi.utwente.nl)

Roeland Ordelman
(ordelman@ewi.utwente.nl)

Abstract:

Filled pauses are generally seen as disfluency. Showing that they can play a positive role in computer generated speech is the purpose of this paper. Details about the causes, variation in pronunciation and syntactic location are acquired through a corpus analysis. This knowledge is then used in an experiment conducted to test positive effects of filled pauses in computer generated speech on listeners. The results show they can influence how well a listener remembers what is said after the filled pause as well as the perceived friendliness and honesty of the speaker and the naturalness of the utterance. This research hopes to contribute to the integration of filled pauses into speech enabled applications.

Preface

Computer agents are designed to mimic human behavior. The study of human interaction allows us to enhance the level of realism of these agents. Studying human speech to improve computer synthesized speech is a perfect example of this. Unit selection synthesizers can already produce very realistic sounding speech through use of recordings of actual human speech. However, there is more to spoken language than the sound of the messages: the paralanguage.

The initial assignment which led to this thesis spoke of laughter, crying and other elements of speech used to express an emotion. The choice of what element I would focus on, however, was left up to me. The next step was to transcribe a corpus to gain an idea of what elements I could choose from and how frequently they occurred. Something became very obvious even before the transcription had been completed. The transcription was littered with occurrences of “um” and “uh”. Their prevalence intrigued me as I set out to investigate these so called “filled pauses”. Although technically filled pauses are contained in many English dictionaries as interjections, making them actual words unlike laughter and crying, their role in speech goes beyond that of an interjection. Their classification is the subject of active debate. As these elements are not very common in speech synthesis, whether or not this direction was in accordance with the assignment’s initial premise is a moot point.

The first goal of the research was to discover the locations of these elements, their frequency, the reason for their presence, their function in the dialogue and their phonetic properties. This information was used to in an experiment conducted to achieve the second goal of this research, which is to determine the effects of filled pauses in computer generated speech on listeners.

I’d like to take the time to thank my family who supported me through this sometimes frustrating project, my supervisors who provided me with feedback and counsel, and finally I would like to thank all those who took the time to participate in the experiment.

NOTATION CONVENTIONS

These apply to all examples in this paper.

Symbol	Description
_	silent pause of normal duration (between 200 and 500 msec)
uh	open filled pause
um	closed filled pause
~	major syntactic boundary
~(…)~	major syntactic boundary with an omitted conjunction
…	key parts inside the examples are encompassed by asterisks

- Noise created by rushing air during inhalation is classified as silence.
- Syntactic boundaries will only appear in examples where the segmentation is relevant
- The example captions are followed by the approximate position of the occurrence in the corpus expressed in seconds.

1 Introduction

Human speech differs from contemporary computer generated speech in many ways, from the robotic sound of rule-based synthesis to the sometimes imprecise approximation of new sound transitions in unit selection synthesis. However, the quality of the *sound* is only one of the differences. The synthesis technologies that don't provide the flexibility to output sounds that are not expressible in phonemes can not imitate certain non-lexical speech elements and other noises humans make without explicitly adding these elements to the vocabulary first. Besides the sound quality and the limited vocabularies the generation of novel utterances is different than that of humans. The language models used for this end produce grammatically correct sentences every time. Human speech, on the other hand, is generally not well formed. Humans make mistakes, stall, restart and correct themselves. This difference is referred to as the "written language bias in linguistics" in (O'Connell & Kowal, 2004). It goes without saying that humans are able to communicate through spoken language quite well, regardless of these disfluencies. As such, we can state that transmitting understandable messages and well formedness need not be at odds with each other. Even though current techniques produce understandable speech, much research is still being done to make synthetic speech sound more human. One of the reasons for this research is because human-like speech can stimulate a user emotionally (Nass & Lee, 2001), which has obvious benefits in certain applications of computer generated speech. Part of the research in this field has to do with vocal paralanguage. Non-lexical elements of spoken language are ubiquitous in daily life but uncommon in automated speech synthesis. This is due in part to the systematic faultless way computer speech is generated.

The prelude to the research discussed in this paper was the counting of non lexical elements in a Dutch language corpus. The result of this process was a large number of filled pauses. Filled pauses, in short, can be described as parts of speech that are uttered in order to allocate time to formulate the next part of an utterance, usually in the form of an "uh" or "um". The research presented in this paper focused exclusively on such non-lexical filled pauses. Due to its abundance the filled pause was chosen for further analysis. The goals of this study became to firstly better understand the *nature* and *properties* of the filled pause through analysis of the corpus. Secondly, to find the *functional roles* that filled pauses play in speech. Thirdly, to *confirm* these functional roles in computer generated speech.

In chapter 2 research material on filled pauses is summarized. Their findings suggest that filled pauses are functional part of spoken language dialogs and not simply 'throwaways' that obstruct comprehension. For example, filled pauses can be beneficial to the listener by preparing them for a word with low accessibility (Corley & Hartsuiker, 2003). Filled pauses can also be the result of social interaction, mitigating undesirable effects of a message. Speakers seem to use filled pauses to increase their apparent fluency and depend on the filtering phenomenon in themselves and in listeners to achieve this effect (Rose, 1998, p. 49). The functions of filled pauses identified by studying literature and were used as a basis for the experiment performed as part of this research.

To find out as much as possible about filled pauses and how they appear in the Dutch language, the transcribed corpus had its filled pauses analyzed. The chapter "Data Collection and Analysis" presents the corpus, the methods used for gathering data and a summary of the results. All filled pauses in the corpus were categorized and several parameters noted. The resulting statistics give an idea of the *frequency* and the *phonetic properties* of filled pauses as well as the *causes* for their occurrence. Information about the location and duration of the filled

pauses as they appear in the corpus were carried over and used in an experiment designed to verify the functions of filled pauses in computer generated utterances.

An experiment was conducted through which it could be determined whether synthetic speech containing filled pauses did indeed have the benefits its natural speech counterparts were attributed. The experiment tested the mitigation and word highlighting functions discussed in chapter 2. It also expanded the scope slightly to verify whether the heightened attention attributed to filled pauses is a side effect of any form of signaling or disruption. The design, execution and results of the experiment are discussed in the chapter titled “Experiment”.

2 Utility of Filled Pauses

The filled pause was first formulated as a psycholinguistic term by Maclay and Osgood in 1959 (O'Connell & Kowal, 2004, p. 459). They are generally considered to be an indicator for hesitation and bad preparedness, but they are also believed to be signals to the listener to indicate a delay in speaking (Clark & Fox Tree). Filled pauses have many realizations including *lexicalized* and *non-lexicalized* forms. The former consisting of regular words and the latter being non-lexical element. Researchers have identified the "uh" and "um" forms of unlexicalized filled pauses in the English language to be two distinct types that indicate the start of short and long delays in speaking respectively (Clark & Fox Tree, 2002, p. 82). The Dutch language has counterparts to the English "uh" and "um" (Clark & Fox Tree, 2002, p. 92). The filled pauses after major discourse boundaries appear to be phonetically as well as functionally different from those at weaker boundaries (Swerts, Wichmann & Beun, 1998). Filled pauses have been categorized as interjections. Although the discussion on the nature of these elements is ongoing, only their causes and functions are important to this paper's topic.

Filled pauses can appear throughout an utterance and it is believed that a speaker uses this time to make linguistic decisions. Filled pauses appear at varying levels of syntactic boundaries and it is believed their duration is related to the type and level of the boundary (Rose, 1998, p. 12). The additional time achieved by pausing is the goal of the pause. The additional time could, for example, be used in computer generated speech in much the same way a human would. Text-to-speech (TTS) systems receive a complete sentence which is then transformed to synthetic speech. In this case there is no extra time needed to formulate the sentence because it is given. More complex speech-enabled applications, however, can make novel sentences by using a language model and a meaning representation for the "beliefs" it wants to express. So, to use filled pauses the same way humans do would be possible if the mapping of a meaning representation to a sentence was an *iterative* process. A filled pause could be uttered whenever the system needs more time to find the right words or structure. The causes and motivation for the use of filled pauses in computer speech need not be the same as in humans for their presence to be useful. The added realism is beneficial by itself, but as the following sections will show, there are more reasons that justify adding filled pauses to the vocabulary of a speech enabled application. These sections represent the three categories into which the functions of filled pauses can be put.

2.1 Turn Management

The reason that filled pause are filled instead of silent is generally believed to be so a speakers can hold his/her conversational turn (Rose, 1998, p. 14). Inversely, filled pauses at the end of a turn can also be used to relinquish the turn (O'Connell & Kowal, 2004, p. 460). Speakers will often use filled pauses at the beginning of a turn to take control of the conversation (Rose, 1998, p. 15). All these functions pertain to the management of conversational turns. In a dialogue humans can use many different means to indicate whether they want to *start*, *stop* or *continue* speaking. These include intonation, gestures as well as filled pauses.

2.2 Mitigation

"An adjacency pair is a unit of conversation that contains an exchange of one turn each by two speakers. The turns are functionally related to each other in such a fashion that the first turn requires a certain type or range of types of second turn." (Loos et al., What is an adjacency pair?) When the second turn falls outside the expected/preferred range and the speaker expects

the listener might perceive the message as unpleasant, then the speaker will use more linguistic effort to moderate its negative effect. As such, filled pauses can be used to soften the blow of an objection or refusal (Rose, 1998, p. 17).

Filled pauses can "downplay the introduction of more difficult terminology or phraseology" (Rose, 1998, p. 17). Doing so allows the listener to be an equal partner in a conversation he/she would otherwise possibly withdraw from. This is consistent with the findings in (Brennan & Williams) which state that listeners often take filled pauses as an indication that the speaker is not knowledgeable on a subject.

In (Eakins & Eakins, 1978) it is observed that females tend to use filled pauses much more often than men, which the authors say might be a way to show some non-assertiveness and submissiveness. This is again in line with the studies regarding the "feeling of another's knowing".

All these situations have one thing in common: the speaker wants to elicit a certain reaction or portray themselves in a certain way to the listener. They are all types of rhetoric that speakers likely use unintentionally as such. The filled pauses in these examples mitigate or moderate the impact of what is being said in order to indirectly benefit the speaker. The speaker's goal in using the filled pauses is to be liked or well thought of.

2.3 Word Highlighting

Words can be implanted into a person's memory if his/her attention is peaked before the word is uttered. Filled pauses are believed to have this effect, making them useful for making sure listeners don't forget important words (Corley & Hartsuiker, 2003). This effect is unintended in spontaneous speech, but could be the result of the listener's interpretation of the filled pause as an indicator for a word with low accessibility.

3 Data Collection and Analysis

3.1 The Corpus

As stated in the introduction a corpus was used to count non-lexical speech elements for the purpose of creating a taxonomy of such elements. That same corpus was used to gather data about filled pauses. This was done in part because it was practical since the transcription had already been done and it was readily available. However, the corpus needed to fulfill certain requirements to be considered at all. It needed to have the following properties:

- Clearly audible speech
- Two or more speakers
- A casual setting with free turn taking
- Spontaneous speech (no trained speakers or memorized monologues)

These requirements were used to get a corpus representative for the kind of casual speech that can be extrapolated to the context of a conversational agent.

The corpus chosen was an episode of a television talk show. The show consists of a host asking directed and undirected questions to five guests which then reply and occasionally talk amongst each other. All but one of the six speakers are proficient in the Dutch language. The least proficient one also spoke very little and his speech omitted from further study. The speakers are at a table surrounded by a studio audience.

3.1.1 Speakers

The five speakers from whom data was collected consist of two males and three females. Here follows a short description of the speakers using subjective and informal measurement of certain parameters. The listing mentions the gender, role in the program. The level of assertiveness is a subjective assertion of the speaker's confidence and aggressiveness in taking conversational turns. Where aggressiveness is affected by the number of turns the speaker claims without having a response directly elicited from them as well as how much energy they put into presenting their opinion. The description also includes whether or not the speaker seems anxious when speaking during the course of the entire program. This is mostly judged by the tone of the speaker's voice.

Speaker	Gender	Role	Assertiveness	Emotional state
1	Female	host	moderately assertive	Relaxed
2	Male	guest	considerably assertive	Anxious
3	Female	guest	moderately assertive	Relaxed
4	Female	guest	marginally assertive	Relaxed
5	Male	guest	marginally assertive	Anxious

Table 1 – Description of speakers with subjective parameters of assertiveness and anxiety. Assertiveness indicates how much initiative the speaker takes and anxiety says whether they're nervous or not.

3.2 Method

This chapter discusses the manner in which the data presented in later chapters was collected. The initial transcription was made in the freely available software package Transcriber (http://trans.sourceforge.net). Every occurrence of a filled pause was noted and categorized according to its cause, which was deduced from the syntax, semantics and in one case the accompanying video. The reason for categorizing according to cause rather than function is discussed in chapter 3.2.7. The largest category was chosen as being the most relevant and easily applicable to text-to-speech systems. It was submitted to further analysis noting the following parameters:

- **Time:** Location within corpus in seconds
- **Duration:** Duration of the element in seconds
- **Sound:** Whether the filled pause was open or closed
- **Boundary:** The type of syntactic boundary at which it appeared
- **Clitic:** Whether the pause was a clitic and if so what kind
- **Double:** Whether the element consisted of two filled pauses
- **Time after:** The amount of time before speech is resumed after a filled pause in seconds

The first parameter should be self explanatory. The rest will be explained in more detail here.

3.2.1 Duration

The duration was measured starting at the point which the waveform became visually distinct from background noise and ending at the point where that was no longer the case. In case of clitics, discussed later in chapter 3.3.3, where the volume of the speech did not lower before and/or after the filled pause an arbitrary position where the phonemes belonging to the filled pauses started and ended was chosen. These positions were chosen so that the following or preceding word was then complete and understandable.

3.2.2 Sound

Open and closed filled pauses differ in the way the sound of the element ends. Closed filled pauses end in consonant ‘m’ sound while open filled pauses sustain the vowel sound until the end. The data collection process produced only a very small number of closed elements: a total of 17 closed filled pauses compared to 265 open filled pauses across all speakers. Multiple corpora cited in (Clark & Fox Tree, 2002, p. 81) contain a much larger proportion of closed filled pauses. The London-Lund corpus they referenced, which contains a similar type of communication as the corpus in this study, contains 1793 closed filled pauses (they refer to them as ‘fillers’) compared to 2111 open filled pauses. The study in (Swerts, Wichmann & Beun, 1998) which used a Dutch language corpus reports 111 occurrences of closed filled pauses versus 199 open filled pauses. The discrepancy between this and other studies could be due to a smaller number of long interruptions in speaking and topical boundaries. This would also explain why the host, being responsible for the topical changes also has the highest frequency of closed filled pauses. However, this was not investigated further.

3.2.3 Boundary

Major and minor syntactic boundaries are useful to indicate where a filled pause is either used to create an utterance or for recalling words. The term major syntactic boundary as used in this paper is defined as the beginning and end of the following:

- Interjections (parts of speech used to express emotion and no grammatical connection to the rest of the sentence)
- Dependant and independent clauses ignoring conjunctions in complex sentences (described in more detail below)
- Incomplete fragments of clauses which have been abandoned or interrupted

The definition of clauses shows how these units can be used to delimit meaning.

> *Independent Clause: An independent clause is a group of words that contains a subject and verb and expresses a complete thought. An independent clause is a sentence.*
>
> *Dependent Clause: A dependent clause is a group of words that contains a subject and verb but does not express a complete thought. A dependent clause cannot be a sentence. Often a dependent clause is marked by a dependent marker word.*
> According to (Independent and Dependent Clauses)

Minor syntactic boundaries are all the locations between words that are not major syntactic boundaries.

The following examples from the corpus will hopefully illustrate the above concepts. The major boundaries in the examples are indicated by a "~".

"Ja ~ je komt nu net uit uh Kamp Zeist ~ hè?"

Example 1 – (152.797) Used to illustrate major syntactic boundaries

"Nou ~ ik ik leef wel heel erg met uh met de mensen in Iraq mee ~ hoor ~(want)~ um als als we daar bericht te horen krijgen dat ze daar uh geintimideerd worden ~(en)~ dat het eigenlijk…"

Example 2 – (275.611) Used to illustrate major syntactic boundaries

"~(maar)~ z.. zou jij uh ~ d'r d'r is een ~ uh uh deze poster hing in alle stembureaus ~ he?"

Example 3 – (603.951) Used to illustrate major syntactic boundaries

In order to measure the frequency of filled pauses at different boundary types two additional parameters per speech turn were collected.

- The number of major syntactic segments
- The average number of words preceding a filled pause at a minor syntactic boundary calculated per major syntactic segment

3.2.4 Clitic

When filled pauses occur in the middle of utterances they sometimes become embedded into the preceding or following word without a pause in articulation. Words that fuse with adjacent

words to form a prosodic unit are referred to as an *enclitic* and respectively a *proclitic*. Refer to chapter 3.3.3 for more information.

3.2.5 Double

Cases where multiple filled pauses occur in sequence they were counted as a single pause. The reason being that it appears when a speaker is very "lost" and needs additional time to think of the right wording the pause will be elongated by repeating what's already been said to fill up the time. There are numeral occasions where the filled pause consists of two uh's and only one occasion where the speaker used three.

3.2.6 Time after

Before annotation of all the parameters of filled pauses started it was observed that the amount of time between the end of a filled pause and the resumed speech was very variable. In order to accurately reproduce this phenomenon the duration of the silence after the filled pauses was measured.

This could also help verify or counter the assertion by (Clark & Fox Tree, 2002) that closed filled pauses signaled longer delays in speaking than open filled pauses. No conclusion was made on this topic because of the small number of closed filled pauses in the corpus.

3.2.7 Categorization

Knowing the events or situations that can cause filled pauses to occur let's us estimate where in the sentence creation process a speaker can use such a pause. That is the reason why the data collection process did not categorize filled pauses according to their function. The functions of filled pauses are a result of analysis and are therefore the creation of the analyst and not a conscious decision by the speaker. It is the way filled pauses affect the listener and the conversation as a whole when viewed from a third-person perspective. A single element may serve many purposes and therefore have multiple functions. In fact, functions are not all distinct and some overlap and correlate with each other. The function of filled pauses has already been studied many times before as shown in the chapter "Utility of Filled Pauses". Causes of filled pauses shed light on the underlying mental process.

There is one problem with categorizing filled pauses according to cause rather than function/effect which is a natural consequence of the complex nature of human speech and the limitations of an audio corpus. A single occurrence of a filled pause buys time which can be used for several mental processes and pinpointing with absolute certainty which one of these was the actual trigger for the pause from the corpus alone is for all intents and purposes impossible. The categorization is therefore an approximation. The consequence of this issue is possible faulty categorization of certain elements, but it does not produce faulty categories. Faulty categorization could affect the way the phonetic properties of the filled pause would be associated with their cause, but the in depth analysis focuses on only one category. The reason for focusing on only the largest category is partly because it is the only one with a reasonable number of elements, but also because it is relevant to our application of filled pauses to computer generated speech.

3.3 Results of the Analysis

The analysis produced data pertaining to the cause of filled pauses as well as their phonetic properties. The following sections discuss this data and present some notable exceptions to what was expected.

3.3.1 Causal categories

Each filled pause in the corpus was categorized according to its most probable cause. The process involved manually creating categories for each filled pause and merging these categories once the similarities became more apparent. The resulting categories are given below.

Cause	Percentage of total	Numbers
Pauses to recall memorized speech	1,15%	3
Pauses to recall facts	5,73%	15
Pauses to select wording	83,97%	220
Pauses for self-correction	8,02%	21
Pauses for reading	0,76%	2
Pauses to take a turn	0,38%	1

Table 2 – Causal category frequencies and actual number of occurrences.

Many of these causes are related to memory. The most notable difference between them is the type of information that is being recalled. Each of these categories will be explained and discussed in the following sections focusing mainly on examples from the corpus. The first category presented below, is not in the above table. From its description it will become clear *why* it's not there.

Pauses for reasoning

Filled pauses give the speaker time to do many things. One of these things is reasoning and deduction. Due to the fact that that process is very hard to distinguish from the creation of a spontaneous utterance a separate category was not created. If recognizing them were easy the pauses for reasoning would still be uncommon in this corpus, because the questions the guests were asked were relatively simple and did not seem to require any deep thought. This is not to say that there was no reasoning occurring. Obviously the speakers are constantly reasoning and thinking, but this ever-present thought did not seem to be the direct cause of any filled pauses.

The following is an example of a filled pause whose cause is unclear but appears at a point where the speaker is obviously engaged in thought. The speaker changes what she is about to say because she wants to address another point first. This indicates that the speaker has come to the conclusion mid-sentence that the current dialogue strategy is not efficient in achieving her goal. The way the first clause is abandoned the speaker is not correcting herself because she hasn't really made a grammatical or enunciation mistake, but rather she's changing the current focus and side-tracks for a little bit in order to then proceed to asking a question.

*"maar z zou jij *uh* de d'r is een uh uh deze poster hing in alle stembureas, he?"*

Example 4 – (605.0) Filled pause followed by changed focus

Pause before reciting canned speech

There are a few cases in the corpus where the speaker recites a piece of text from memory. This is different from regular spontaneous speech because the speaker does not create the utterance at the time it is being said. The pause used to recall these sentences is therefore different from those found in spontaneous speech.

There is an occurrence before the first sentence of the program. After the host greets the crowd she introduces the topic of the day but not before uttering a brief pause. It is a very short pause that is easy to miss. Since she is a professional talk show host we can assume that she has prepared for this show and knows this introduction sentence well. This is supported by the fact that there are no pauses or other forms of hesitation in the first sentence. As a result, the presence of the pause can not be explained by low accessibility of the following words or by the effort needed to create the sentence. The reason for this pause seems to be that the speaker had to recall the sentence as a whole.

Constructions like expressions and figures of speech consisting of multiple words can be recalled as a whole. This seems to be different from regular spontaneous speech. The filled pauses before canned speech indicate a search for a string of words with a specific order that has been memorized beforehand. One may argue that spontaneous speech is nothing more than concatenation of memorized speech. After all, conjugations like “ik ben”, “jij bent” and “zij zijn” are used so often that the combination of the two words form a unit of sorts. Wouldn’t recalling such a unit is therefore akin to recalling a memorized sentence? The answer it seems is ‘no’. The difference between recalling canned speech and the construction of spontaneous utterances can be shown through means of an example. When reciting the words of a song you may be aware of their meaning, but the *awareness* does not necessarily *precede* verbalization as is the case during spontaneous speech. It is like the fact that is made obvious to Alice while in Wonderland, which is that “meaning what you say” is not the same as “saying what you mean”. While a speaker is be able to make such a distinction, a listener may not recognize when the speaker is quoting from a foreign source.

Pause to remember facts

The corpus contained several pauses that were followed by names, facts or numbers which were somewhat isolated from the rest of the utterance. These facts about the world are the product of a different mental process than the creation of an utterance.

In the following example the speaker lists a number of locations and pauses before the last one. She needs some time to remember this last element of the group, namely “Zwolle”. Shortly after this another short pause is used to recall the name of the person that is being addressed.

*“Je bent in Amsterdam Rotterdam en *uh in Zwolle* geweest want je hebt ook meegeholpen uh met de organisatie doe je nogsteeds _ *uh **NAME**”*

Example 5 – (108.9) Showing filled pauses when recalling a fact. Name omitted from example for privacy reasons.

Pause to select wording

Utterances are usually “conceived and composed by their speakers even as they are spoken” (Mehta & Cutler). Thoughts that the speaker wants to express need to be converted to words and word sequences that help convey the thoughts and concepts that he/she wishes to communicate to the listener. The fact that the pauses often occur in the middle of the utterances

indicate that this conversion happens even while the speaker is talking. The mistakes show that this process isn't perfect and is influenced by other factors besides the concepts to be communicated and the meaning of words, factors such as the perceived effects of previous words and the supposed effects of future words on the listener's and the speaker's emotional states. The fact that different people will often create different sentences to express the same concepts hints to this as well

A large number of the filled pauses observed in the corpus appear in the middle of phrases. These pauses afford the speaker time to make linguistic decisions. They help the speaker determine what words and constructs will be used. The utterance in the next example is preceded by a pause and it contains three others. The parts of the utterance that are separated by the pauses don't make it obvious why the pauses are there. The explanation that accounts for this is that the speaker knows what she wants to say but is trying to figure out how to say it. Instead of having a very long pause while she figures out the entire sentence, she prefers to utter the parts as she creates them. This is normal behavior in spontaneous spoken language, the exception being trained public speakers who make a conscious effort to minimize the use of filled pauses. The placement of the pauses seems to facilitate the process of piece-by-piece sentence construction. One could interpret the pauses as audible delimiters of sentence units/modules which connect to create the eventual utterance. Humans can create grammatically correct sentences without much effort as is proven by people every day.

*"*uh* Jij bent *um* een van die *uh* vele jonge Irakezen in Nederland _ die *uh* betrokken zijn geweest bij de…"*

Example 6 – (42.0) Several filled pauses used for word selecting. The cause of the first filled pause is debatable.

The first pause happens before the sentence starts. Because of this it is entirely possible that the host is still figuring out what to say and not just how to say it, giving it a different cause. However, these two activities probably happen at the same time or in quick succession. Many elements that serve multiple purposes like this occur in the corpus.

Pauses for self-correction

When correcting one's self a filled pause can be used to void whatever was previously said. The scope of the correction becomes obvious once the correction is given. The corrections happen during spontaneous speech and are very similar to the pauses used to select wording. The difference being that when correcting one's self the right wording has been found after the speaker has already partially committed to saying something else.

In this example the speaker corrects herself replacing the word "wat" with the word "veel".

*"Nu is het wat *uh* veel rustiger"*

Example 7 – (173.7) The speaker corrects herself and replaces the word "wat" with "veel".

Some corrections in the corpus happen because the speaker has mispronounced a word. This is a different kind of mistake than the one above. There is no reason to suspect that the pause has any other motivation than to reattempt pronouncing the word. Not all corrections use filled pauses, especially when a mistake is a mispronunciation the speaker often just tries again without a pause or hesitation.

The young lady in the next example is trying to say something but corrects herself twice. She is trying to choose between two different expressions using the words “vergelijking” and “voorbeeld”, but unfortunately she commits to speaking before having made up her mind and therefore must correct herself.

*“nee want bijvoorbeeld laten we *um* voorbeeld heel ve *um* ja, een vergelijking, wat er gebeurd net een tijdje terug in Azië”*

Example 8 – (887.8) Two cases of self correction in quick succession.

Pause for reading

Early on in the corpus there are a few cases of the host needing to look at a piece of paper, which can be seen quite clearly on the video. The pause coincides with the look and suggests that she is reading during this time. She doesn’t read the entire sentence off of the paper and used it only as a cue. It is similar to actors who forget their lines. A person can be prompted only the beginning of a memorized sentence and recall the rest of it.

*“_ um _ uh **NAME**”*

Example 9 – (39.5) A long filled pause that coincides with reading by the host, followed by the name of the person to whom she’s directing her next question. Name omitted for privacy reasons.

In Example 9 you can see a rather long pause consisting of an ‘um’, some silence and an ‘uh’. What follows after the pause is the name of the person to whom the host is directing her conversation. She is making it clear that the following questions are for him. During the pause the host looks at her papers. She could be looking for either the question she is about to ask or the person’s name. The latter seems more likely, because the question she asks appears to be spontaneous and after she says the name she stops looking at the papers. There is no way to be certain, however.

Both these pauses mentioned here are paired with reading. They are searches on paper rather than searches of the speaker’s memory. This causal category can be generalized as pauses that are used to delay speaking until a non-speech related action has completed.

Pause for turn management

A filled pause can be used to evaluate a situation and to decide whether or not it is ok to speak. When a question is asked to the guests as a group the person who answers starts with and ‘uh’ followed by an answer. The question was who had the most ink on their finger. The comparison of fingers had already taken place and the speaker did not need to look to the other guests to know the answer. Assuming that the fact that he was the one with the most ink on his finger was still fresh in his memory, we can safely assume that the ‘uh’ was not mainly used in recalling a fact. The answer consisted of a single word: ‘ik’. This would suggest that the ‘uh’ was not used while constructing a sentence either. Taking these facts into consideration we find that the ‘uh’ was probably meant to indicate to the others that the speaker was going to take his turn and start speaking.

3.3.2 Frequency of Filled Pauses

Data was collected on the frequency of filled pauses in the corpus in order to compare the numbers across all speakers to give a good idea of how much or how little the frequency varies.

Frequency of filled pauses

The following table gives an indication of the frequency of filled pauses for each speaker. The total number of uttered words includes filled pauses.

Speaker	Pauses	Wordcount	Pauses per 100
1	80	1716	4,66%
2	31	1132	2,74%
3	29	587	4,94%
4	56	610	9,18%
5	86	1091	7,88%

Table 3 – Filled pause frequency

Based on these numbers it seems obvious to conclude the presence of a relationship between assertiveness and frequency of filled pauses. Such conclusion may be premature due to the small number of speakers and subjective nature of judging aspects of a person's personality. However, this relationship is in line with what was found in (Eakins & Eakins), although the relationship between frequency of filled pauses and gender of the speaker is not very pronounced in this corpus.

The frequency of filled pauses varies greatly across the speakers, but the distribution across the two boundary types does not. The proportion of the total number of filled pauses that appear at minor syntactic boundaries averages to 62% across all speakers. The greatest deviations from this average are 51% and 70% as can be seen in Figure 1.

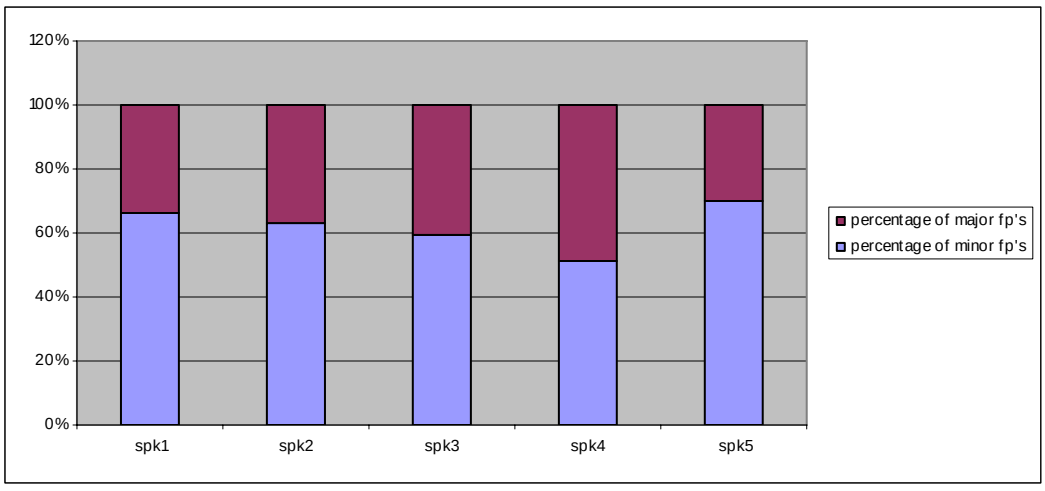


Figure 1 – distribution of filled pauses among the two boundary types for all 5 speakers. On average 62% of all filled pauses occur at minor syntactic boundaries.

3.3.3 Phonetic properties

Clitics

In (Clark & Fox Tree, 2002) it was found that filled pauses always appear encliticized (leaning on the previous word) and never procliticized (leaning on the following word). In contrast, the corpus in this study contains several procliticized elements. In Figure 2 a waveform can be seen of one such procliticized filled pause.

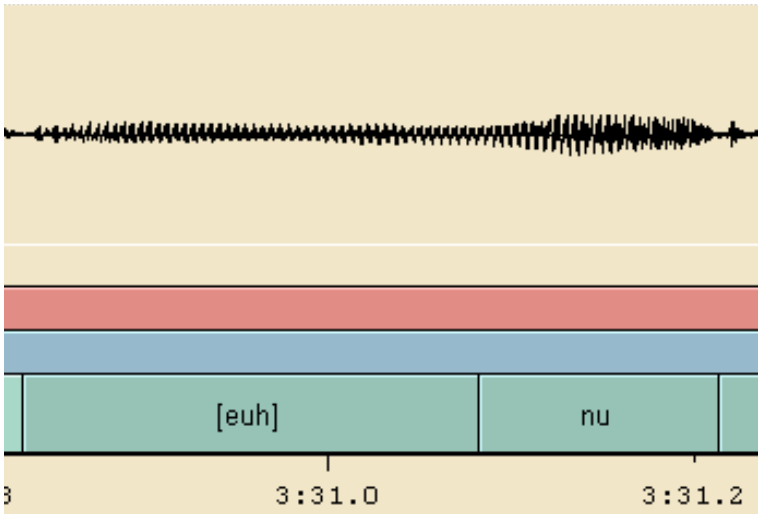


Figure 2 – procliticized filled pause. There is no silence between the vocalization of the filled pause and the following word. Found at 3:30 of the corpus.

Figure 3 shows an example of an encliticized filled pause. As with both types of clitics the filled pause has become part of an adjacent word forming a single prosodic unit. In this particular example the same ‘uh’ sound we recognize from other filled pauses is present here but as part of the word ‘de’, making it sound like a long ‘duh’. The duration of the ‘de’ is much longer than a normal occurrence which is why it was transcribed as a pause. Not all cases of enclitical pauses are as clear as this one. Sometimes a ‘de’ is stretched without it being clear whether or not it is a pause. Other occurrences of an embedded ‘uh’ are clearer to see because the last sound of the previous word is not the same as an ‘uh’.

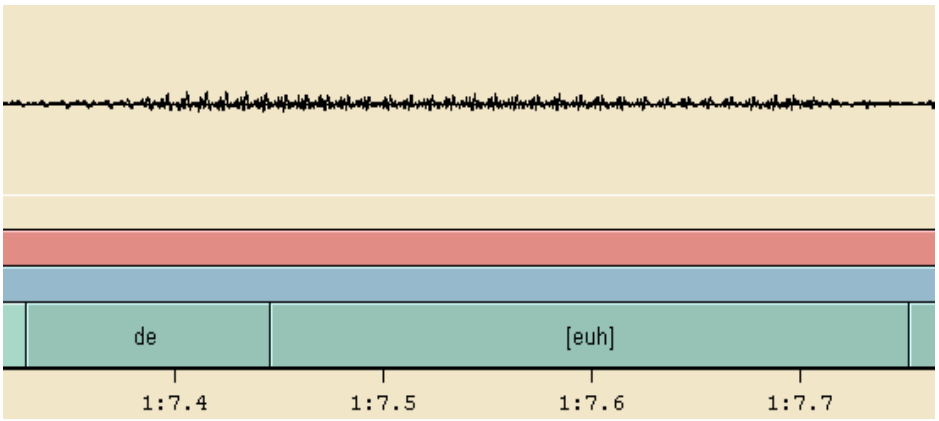


Figure 3 – The waveform of an enclitic filled pause. The transition from one to the other is not clear because the vowel sound of the word "de" has is the same as the following filled pause.

The corpus gave a fairly good indication of a relationship between boundary type and cliticized elements. Four of the five speakers had a majority of the cliticized elements at minor syntactic breaks and the one exception had only one less cliticized element at minor breaks than major breaks. No relation was found between the type of clitic and the boundary type nor is there any consistency in the prevalence of one type of clitic over the other.

Varying pronunciation

The corpus shows slightly different phonetic variations among the filled pauses even within the same categories. This chapter discusses two examples of such variation.

In the excerpt in Figure 4 both parts of the double filled pause end with a 'b' sound, which could be interpreted by a listener as a sign that the speaker's next word will start with a 'b'. However, after repeating the 'uhb' the speaker follows up with the word "was" as can be seen below. This is not in accordance with the expectation the 'uhb' created. Perhaps the speaker changed her mind. It is not obvious what caused the speaker to pronounce the 'uh' differently here.

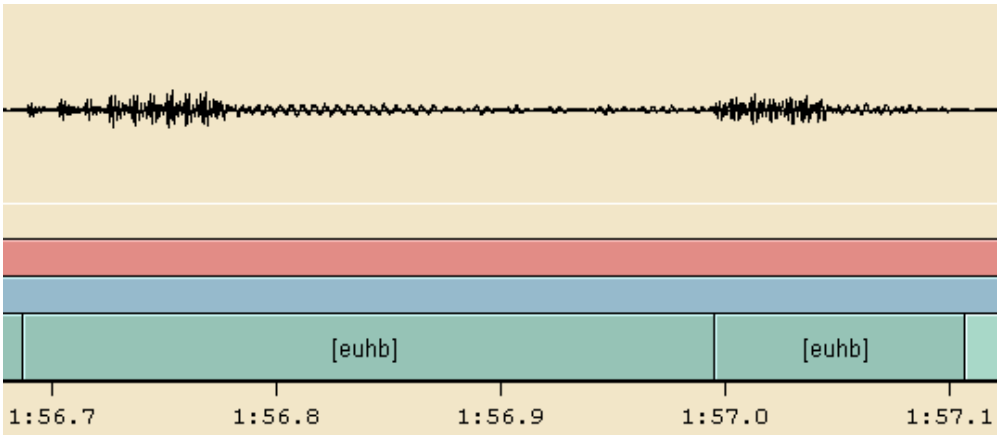


Figure 4 – An exceptional occurrence of a double filled pause ending in a consonant 'b' sound.

In Figure 5 you can see a short subtle pause that is easy to miss when listening to the actual

audio. It is a raspy sound that is different from the majority of filled pauses. This type of element is so soft that one could argue that they aren't filled at all. However there is an audible sound and the effect on the speech is like that of the other filled pauses inside utterances. The effect on the surrounding utterance includes the intonation and tempo of articulation. There are several instances of soft and short pauses that sound unlike the rest.

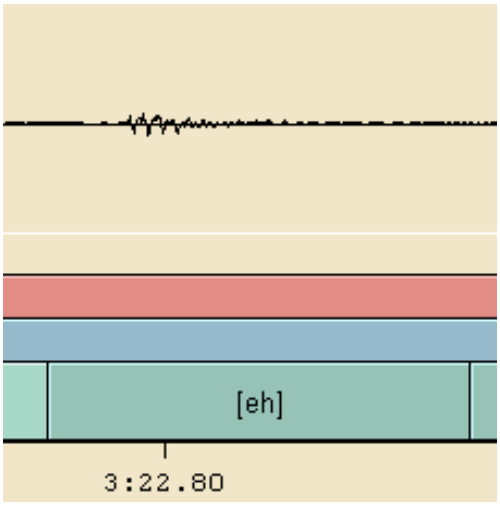


Figure 5 – a very short filled pause that sounds different than most filled pauses. The duration of the marked area is only 0.04 seconds.

Duration

The average duration of the analyzed elements comes to 0,318 seconds. Obviously, this number alone does not reflect the full spectrum. There were many extremely short or very long filled pauses as evident from the example in Figure 5. Minor and major syntactic boundaries are locations where different kinds and amounts of information are processed so one would expect that difference to be reflected in the duration of pauses at each type of boundary. The following chart shows the two types of filled pauses for each speaker.

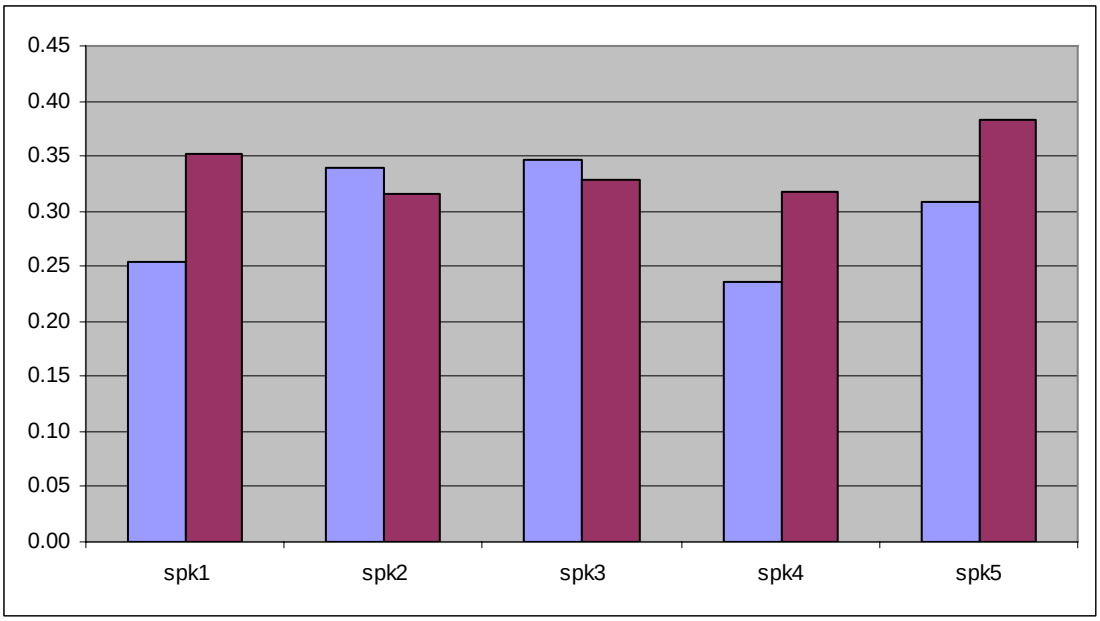


Figure 6 – average duration of filled pauses per speaker. The first column for each speaker is the average duration of pauses at minor breaks and the second is the average duration of filled pauses at major breaks.

Across the speakers filled pauses at major syntactic breaks are not consistently shorter or longer than those at minor breaks. What can be seen is that the duration of filled pauses at major breaks varies less across speakers than those at minor breaks. The following chart shows the amount of variance in duration across the five speakers.

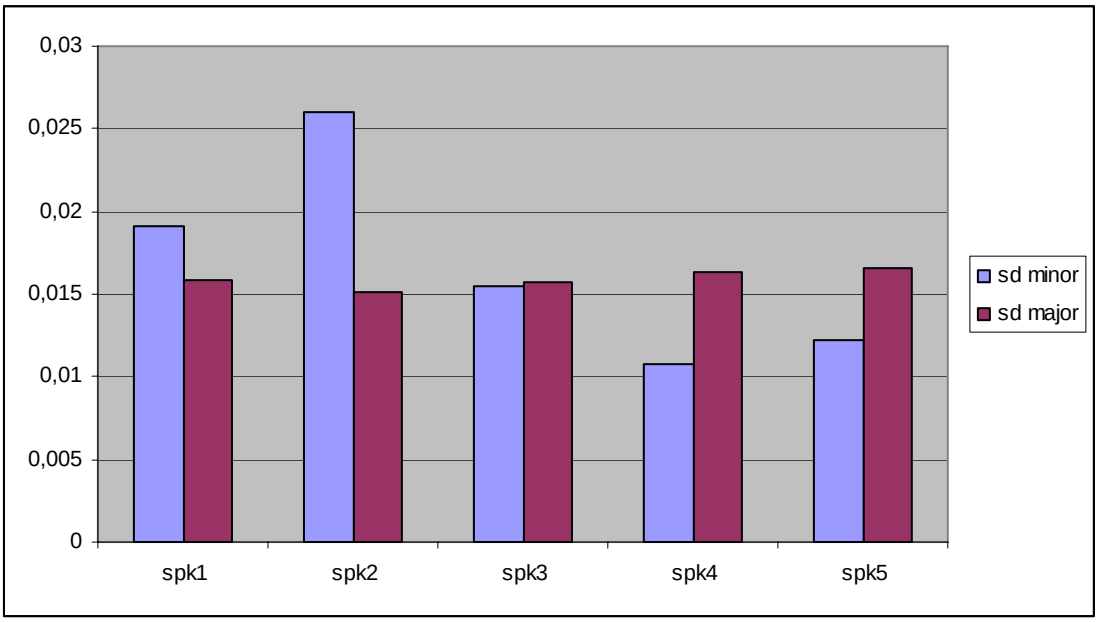


Figure 7 – standard deviation of FP duration. First column per speaker is the SD of filled pauses at minor breaks and the second of pauses at major breaks.

This table shows that the standard deviation of the duration for filled pauses at major syntactic boundaries is very consistent across all speakers. In (Clark & Fox Tree, 2002) it is suggested that filled pauses are words which are planned for. If filled pauses at major boundaries are standard practice for a specific speaker, then you would expect their duration to be consistent. Conversely, unexpected/emergency filled pauses for remembering words mid-sentence will last for as long as needed are very unpredictable because of that. The standard deviation of filled pauses at minor breaks is an indicator for how speakers deal with unexpected pauses.

When comparing the values in Figure 7 to the frequency of filled pauses of the analyzed category shown in Figure 8 you might notice a relationship. The speakers with a large standard deviation in duration at minor breaks are also speakers who used fewer filled pauses.

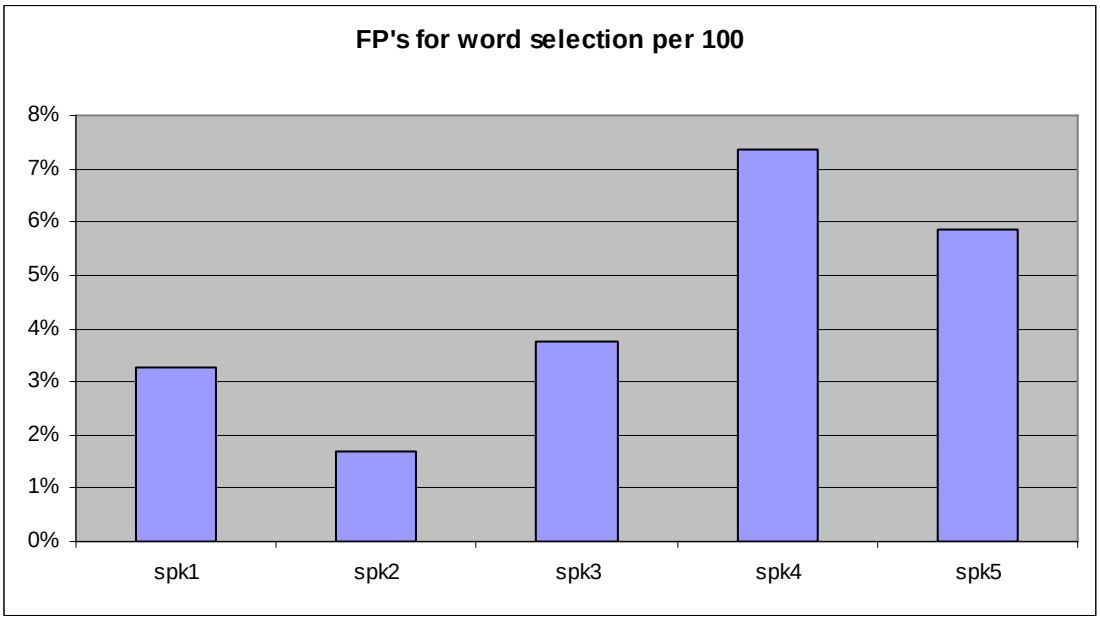


Figure 8 – Frequencies of select wording FP's in percentage of spoken words per speaker.

The relationship was a surprise and its consistency makes it seem more than coincidence. This phenomenon can be explained by the assertion that speakers who use filled pause often also plan for them more than speakers that use fewer filled pauses. If this is true then the 'planned' elements would exhibit commonalities including duration.

The duration of the silence after filled pauses as well as how many filled pauses actually had silence after is shown in Table 4.

Speaker	Avg. duration	Num. FP's with silence (total)
1	0.07	19 (56)
2	0.13	8 (19)
3	0.10	10 (22)
4	0.06	13 (45)
5	0.07	23 (64)

Table 4 – duration of silence after filled pauses. The second column is the average duration of the silence after the analyzed filled pauses in seconds. The third column shows how many elements actually had any silence after. The number in brackets is how many elements this parameter was taken from.

4 Experiment: Effects of Filled Pauses

4.1 Introduction

The corpus analysis resulted in some interesting information about the frequency of occurrences of filled pauses and their phonetic aspects. This information answers certain question with regards to natural language, but doesn't say anything about its relevance to the field of speech enabled agents and applications. The effects of filled pauses on listeners are still being studied, but what is currently known can be put to the test using synthetic speech in the form of an experiment. This form of experimentation attempts to bridge the gap of knowledge between the psycholinguistic theory and spoken language interaction between computers and humans.

Before the experiment came to be, an application was developed which was meant to facilitate this type of experimentation while at the same time performing some sort of practical task. This application was designed and implemented to be a speech enabled appointment manager called PAM (pausing appointment manager). The application was created but abandoned in favor of the experiment because of the greater flexibility it provided. PAM dealt mainly with numbers (times and dates) which meant results of testing would not be as generalized. The domain covered by PAM required unintuitive additions in order to test any effect other than word highlighting. The on-the-fly generation of speech required compromises to the quality of the sound. Though the TTS engine used in both PAM and the experiment is the same, the pre-generated nature of the audio of the experiment application gives a greater amount of control. PAM, due to its shortcomings did not provide an adequate platform for testing the multiple functions of filled pauses that are tested in this experiment.

The following sub-chapters discuss the theoretic aspects of designing the experiment, detailed information about the experiments implementation and how the experiment was conducted.

4.1.1 Problem Area

The experiment is intended to find evidence for the claim that filled pauses in computer generated speech play a positive role. The chapter "Utility of Filled Pauses" shows some positive side-effects filled pauses can have in a conversation. Among these is the heightened attention of test subjects in the experiments conducted in (Corley & Hartsuiker, 2003) as well as the mitigation of negative effects in adjacency pairs as presented in (Rose, 1998, p. 17). The former is an objective property affecting performance and the latter affects a subjective impression of friendliness. These two effects may or may not be limited to spontaneous human speech. This experiment is meant to explore this question. In addition to extending the two effects to the domain of synthetic speech, this experiment tries to give an indication of user perception of computer generated speech with filled pauses. The outcome of these tests can be interpreted as evidence for or against filled pauses in speech-enabled applications.

4.1.2 Causal Relationship

Firstly, this experiment examines the heightened attention triggered by filled pauses during synthetic speech and its effect on the listener's ability to accurately recall the elements following the filled pause. In addition to this, the experiment tests whether the heightened attention is unique to unlexicalized filled pauses or if it can also be achieved with an artificial non-verbal signal. Secondly, this experiment examines the effect the presence of filled pauses

has perceived differences in friendliness of replies in adjacency pairs in both natural and synthetic speech. Lastly, this experiment compares the preference of listeners for computer generated speech with and without filled pauses in several categories.

4.1.3 Hypothesis

The research questions and hypotheses for this research are:

RQ1: Do filled pauses in front of a word affect how well it is stored in memory?
Hypothesis 1: A filled pause in front of a word causes heightened attention which improves the chance that a listener will be able to recall it.

RQ2: Do filled pauses alone affect the perceived friendliness of a dispreferred answer in an adjacency pair?
Hypothesis 2: Filled pauses in the second part of adjacency pairs mitigate their negative effect, making them seem friendlier.

RQ3: Is heightened attention triggered by filled pauses because of their special nature or would artificial signaling be equally effective?
Hypothesis 3: Synthetic speech with filled pauses has a stronger positive effect on a listener's performance when recalling words than an artificial acoustic signal.

RQ4: Do listeners prefer synthetic speech with filled pauses or without?
Hypothesis 4: Listeners, on average, will not show any preference between speech with filled pauses and speech without filled pauses.

4.1.4 Computational Model

The corpus analysis yielded a lot of information which was not directly applicable to the experiment. The goal of the experiment is to confirm certain functions of filled pauses. In order to do so the realization of filled pauses is controlled by knowledge gained from the corpus analysis. However, the variation in phonetic properties was not controlled because the relationship between the realization and the location of the filled pause was still unclear after the corpus analysis.

The *average duration* of the filled pauses was used as a baseline in all experiment materials and was tweaked from there until it was subjectively judged to sound natural. The corpus analysis did not include research into the intonation of the filled pauses and its relationship to the intonation of adjacent words. However, during the creation of the experiment materials outside input was used to fine tune this parameter into something that pre-test subjects would find natural.

The location of filled pauses with regards to the syntactic boundary was irrelevant to the filled pauses used in the experiment because of its unknown effect on the actual sound. However, the materials used in the exercises 2 and 3 used sentences whose structure, including the location of the filled pause, was taken from sentences contained in the corpus.

The chapters devoted to the design of the individual exercises contain sections on the materials used. Those sections explain the finer details of how the corpus analysis relates to the way the filled pauses were made.

4.2 Method

4.2.1 Technical Details of the Experiment Application

The experiment is performed entirely on PC's using a Java application. The experiment has been divided into three parts. The first experiment is designed to test hypothesis 1, the second part tests hypothesis 2 and the final part is designed to establish a preference on different categories to test hypothesis 4. A variation of the materials in part 1 tests hypothesis 3.

The invitation to participate was sent through direct emails and through instant messaging indicating only a request to participate, the approximate duration and a URL to the starting page of the exercise. The starting page contained information on what was expected from users, the system requirements for running the application, the security risks for running the application and the privacy policy applicable to their personal information. The exact text is available in Appendix B. It is worth noting that the Java application is started through the browser and is loaded using Java Web Start. This is to make the installation process as simple as possible without having to make the experiment an applet built into the web-page. Doing this also ensures a certain level of security which is explained on the starting page. Java version 1.4 is required to run the application, which was consciously chosen as it is considered to be the lowest common denominator across most platforms. Code was tested on both JVM 1.6 and 1.4 before the experiment was officially started.

The experiment application is designed to be user friendly, simple and consistent in its functionality. A sample image of the GUI can be seen in Figure 9. It has instructions for each step of the experiment. When input is required only those fields that should be used are activated, when something is not used it is disabled. Before the first exercise the application gives the subject a chance to listen to a piece of synthetic speech similar to what they will hear during the exercises. This is an opportunity for the subject to adjust his/her speaker volume and to get accustomed to the sound of the computer generated speech. After the exercises have been completed the application stores its data on a central database.

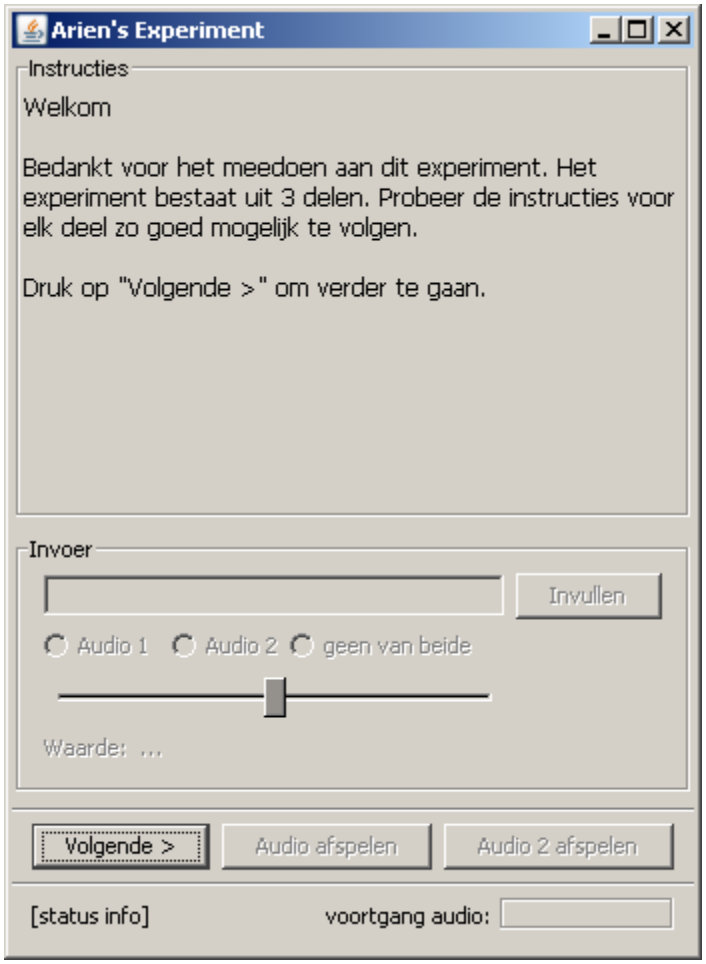


Figure 9 – GUI of the experiment application.

The database is a MySQL 3.22 server located at the University Twente. The relational database's structure is presented in Appendix A. The database is contacted 3 times during the experiment. Once to see if the email address input by the user is unique to the database. A second time to count the number of times each version of the experiment was performed up to that point to help determine which version of the experiment should be used for the next one. The last time it is contacted the data from the completed experiment is stored in the database. The database does not use any keys in order to simplify the amount of error checking on the client side. The client does however use part of Java Web Start's basic service package to store experiment data locally if for whatever reason the database could not be contacted during the data storage phase. This service makes the application robust and able to deal with a sudden connection loss. The data stored locally can be submitted the next time the application is run. The user is informed about this when the first attempt fails and is asked to retry the submittal once the connection can be established. The application will refuse to run the experiment until the previous data is submitted. Once this has been done the application can be restarted and will function in the same way as it did the first time.

All audio content was generated using the Fluency text-to-speech system. This software package enables the creation of customized utterances with adjustable intonation. Several rounds of pre-testing were conducted to refine the utterances, content and layout of the application. The utterances were generated beforehand as opposed to generated during run-

time. This works more reliably and allows for more precise control over the phoneme content, tempo and intonation curve. The synthesis system used by Fluency is a diphone synthesizer. As a result the audio is easily recognized as being synthetic. The effect of filled pauses in synthetic speech need to be measured in speech that the listener can clearly distinguish from natural speech, because the effects analyzed are obviously the effects on the listener and not on the speech itself. An additional requirement to the TTS system is precise control over the generation of a filled pause. The generated audio was composed and compressed using the free open source sound editor software Audacity.

The application was developed using the NetBeans IDE and its integrated GUI authoring mechanism Matisse. The generated GUI uses code which is only integrated into the JRE in version 1.6. Consequently, for increased backwards compatibility the required extensions were distributed along with the application. The application also depended on other external objects, which were also packaged into Java archives and distributed accordingly. These packages include:

- MySQL Connector/J: *MySQL AB's JDBC Driver for MySQL by MySQL AB*
- MP3SPI: *Service provider Interface that adds MP3 support for JavaSound by JavaZOOM*
- two small packages required by MP3SPI: *jl.1.0.jar* and *tritonus_share.jar*

The next section explains the subject information gathering. The sections after that cover the design, materials and procedures of the three sub-divisions/exercises that compose the whole experiment. This is then followed by a section on user selection.

4.2.2 Subject Information

Test subjects are asked to fill in some personal information before the experiment starts. An image of the screen can be seen in Figure 10. The data includes:

- E-mail address
- Age (0-18, 18-30, 31-40, 41-60, 60+)
- Gender
- Whether Dutch is their native or second language
- Experience with Computer Generated Speech (none/some/a lot)

Figure 10 – User info prompt. Displayed at the start of the experiment application.

All user data is intended to discover possible relationships between experiment results and a certain demographic. The e-mail address however is intended to identify each individual subject and at the same time function as a way to prevent the completion of the experiment multiple times by the same participant. This is in no way fool proof and it is present simply as a deterrent. If subjects wanted to fool the system they could, but care was taken not to introduce any elements that could encourage them to do so. For instance, the idea of having a lottery among the users to attract more participants was not implemented. The possible reward could encourage some people to commit fraud. The original intention of the e-mail address information was to give the subjects the option of choosing to be informed about the meaning and results of the experiment. However, this hasn't implemented. The email addresses are checked for duplicates in the database. A syntax check is done on the email address. This check complies for the most part with the appropriate RFC specification. The code was originally written by Les Hazlewood who is credited here as well as in the application source code comments.

The web page from which the application is launched instructs users as to the privacy of their information. The policy is quite simple. None of their information will be sold or given to any company or individual for any purpose other than the statistical analysis of this experiment's results.

4.2.3 Exercise 1: Word Highlighting

Design

The first experiment is a free recall experiment, a classic device for measuring memory. The serial position curve present in the results of such an experiment show how subjects have a harder time recalling the items in the middle of the list than those at the start and the end. This is due to the so called primacy and recency effects. These terms stem from the theory that human memory consists of two separate stores: short term memory and long term memory.

According to this theory, when memorizing a sequence of items the most recent items are stored in short term memory and are easily recalled (recency effect). Items at the beginning of the list were rehearsed enough to be put in long term memory (primacy effect). Words in the middle are said to decay because of the lack of rehearsal time and the limited capacity of the short term memory (Murdock, 1962). Although some consider this theory to be an outdated and overly simplified model of actual memory, the serial position curve does exist. Knowing this, words in the middle of the list follow a filled pause in order to affect how frequently they're recalled. The effect on the serial position curve as well as the specific words give an indication of what influence the filled pauses have.

There exist three versions of the word lists of which two are slight modifications of the original. The first and "original" version contains the word list with a steady rate and no additions. Another version contains filled pauses in front of specific words among the first six of the twelve items. The third version is uses a non-verbal acoustic signal instead of a filled pause for reasons discussed in the introduction (see Hypothesis 3). The aforementioned versions of each list will be mixed so that subjects do not realize the point of the exercise and focus only on those words preceded by a signal.

Materials

Participants are presented with 6 lists of 12 nouns as shown in Table 6. The nouns in each list are common Dutch words. The first three lists consist of related words and the last three of unrelated or weakly related words. This division was made in order to give varying difficulty so that recall rates for a single user would not be consistently high or low.

The presentation rate for the original unmodified version is 1.5 seconds per word. Due to the length of the filled pauses and the acoustic signal the words following the signals had to be moved forward in time the length of the signal's duration. As a result the modified versions have a slightly lower presentation rate. A lower presentation rate increases the amount of time a subject has to rehearse words he/she has heard. If this happens words before the signal could be recollected more frequently. The point of interest for this exercise is the recall rate of the selected words optionally preceded by a filled pause or tone signal. The additional time to rehearse list items should, in theory, only affect items preceding the signal. Taking these two facts into account, the change in presentation rate will at most affect one signaled item (item preceded by a signal).

The signaled items fulfill certain requirements. Before choosing the signaled items it was decided that no two adjacent words should be signaled. This separation is present to avoid too much temporization around a single word. If a word was surrounded by filled pauses it would not be possible to infer if the effect on performance (if any) was the result of the filled pause signaling to the user to focus more or if it was due to the longer time to rehearse the word provided by the following filled pause. This is similar to the issue of presentation rate discussed in the previous paragraph. The second requirement for signaled items was briefly mentioned in the "Design" section for this exercise. The requirement has to do with the position of the items in the list. The experiments in (Glanzer & Cunitz, 1966) show how the primacy and recency effects function. These are the two factors are seen as the cause of the serial position curve in the results of free recall exercises. This curve shows that list items in the middle, on average, have a lower recall rate than those at the start (primacy) and the end of lists (recency). The two requirements and the length of the lists inspired the choice to have two signaled items per list.

The distribution of experiment material occurs in such a way as to ensure each version is completed an equal number of times. Each participant will receive an equal number of unmodified lists, lists with filled pauses and lists with a tone signal. The multiple versions of each list will be combined as shown in the following table.

List \ Combination	I	II	III
List 1	O	FP	SI
List 2	FP	SI	O
List 3	SI	O	FP
List 4	O	FP	SI
List 5	SI	O	FP
List 6	FP	SI	O

Table 5 – mapping of the three versions of each list to the three combinations. "O" indicates the original version. "FP" indicates the version with filled pause signals. "SI" indicates the version with artificial acoustic signals.

The combinations labeled with roman numerals in Table 5 are distributed pseudo-randomly. Each time the experiment is performed the application checks to see how many of each combination have been completed and decides to give the participant the combination that has been performed the least. If two or all three combinations are tied for the least number of completed experiments, then one of those combinations is selected randomly.

The filled pause signal is generated using the "shwa" phoneme. All utterances in the experiment were created using the Speech Editor program, which is part of the Fluency software package. Inserting a "shwa" phoneme into the utterance by manually editing the appropriate file produces an utterance that retains the intonation of the original, but only slightly stretched out over time. This produced very realistic results most of the time. In those cases where this was deemed not to be the case the intonation curve was adjusted.

The schwa is the most common unstressed vowel. It appears often as the result of vowel reduction (Jurafsky & Martin, 2000, p. 62). The schwa vowel is neutral in its articulation and as a result sits in the middle of the vowel chart shown below in Figure 11 represented by the symbol **ə**.

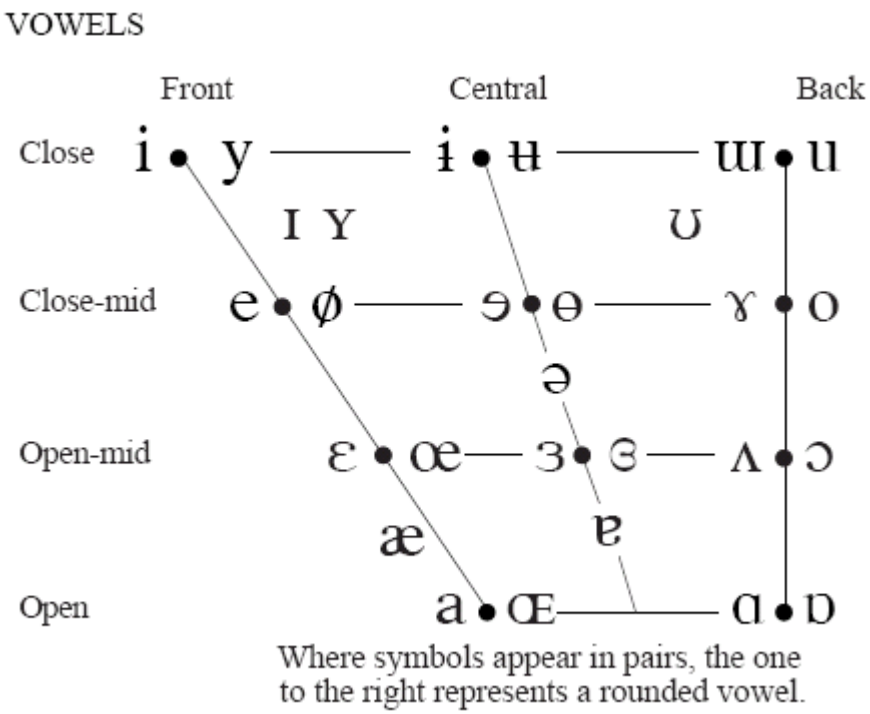


Figure 11 – vowel chart. The labels on the two axes roughly correlate to the position of the tongue and the mouth during pronunciation. The schwa symbol (ə) sits in the middle. (http://www.arts.gla.ac.uk/ipa/vowels.html)

The position of the articulatory organs while uttering a filled pause is defined by the preceding and following sounds. Although clitic occurrences of filled pauses do have a transition between preceding and following words, the middle of the filled pause can still be accurately approximated with the schwa. There are slight variations of the schwa phoneme which are unavoidable due to the many factors that influence its pronunciation. Even the sound of non-clitic occurrences of filled pauses is affected by such foreign factors as is apparent from its variance. Although the vocalization of filled pauses or any word is not independent from the adjacent sounds, informal testing performed before the creation of the experiment materials has shown that the central schwa is an appropriate generalization. The intonation is specified using the "pitch nodes" in the Speech Editor to go from neutral to whatever intonation the word would normally have been given by the software. Subjective judgment was used for choosing the specific location of these nodes. The duration of the filled pause signal was set to approximately 0.3 milliseconds. A value based on the data from the corpus analysis.

The non-voice signal in the second modified version (third version overall) of the lists is a recording of a tone used in a public announcement system. This was chosen because of how filled pauses similarly announce the start of something. The subjects, as with the filled pauses, are not prepared for the acoustic signaling. This fact was made all too clear during pre-testing when complaints were made with regards to the 200Hz tone that was used at the time. The tone sound made users think there could be something wrong with the audio. This is the main reason why the current signal is more complex and easier to identify as a recording of an actual sound rather than a procedurally generated sound.

List	Words
1	cabriolet, fiets, racewagen, auto, **motor**, vliegtuig, truck, **bus**, rem, stuur, Mercedes, snelweg
2	huis, venster, dak, **kamer**, deur, trap, **stoel**, kast, stoep, tuin, hek, plafond
3	artikel, schrijver, verhaal, boek, **brochure**, gedicht, strip, lied, **roman**, schrift, krant, papier
4	patroon, stal, spel, **draad**, mok, plaat, **raam**, plastic, ventiel, broer, poster, lamp
5	kronkel, kooi, tak, berg, potlood, **rook**, bureau, **scherm**, handvat, legende, paard, winkel
6	pion, kameel, **ruimte**, sneeuw, horloge, **vuist**, slijm, jaar, doorn, afslag, foto, broek

Table 6 – Lists of nouns used in experiment 1. The words in bold are preceded by signals in the appropriate versions of the audio.

Procedure

Subjects are given instructions for the free-recall task. At this point the application will already know which combination of lists to play (see Table 5). After the initial screen with instructions is shown the user can more on to the first list which comes with its own set of instructions. This screen explains what the buttons do and how to input their results. Subjects can either press the input button or the 'enter' button on their keyboard for faster input. The play buttons at the bottom of the screen allow the subject to start the audio clip containing the word list. Each audio clip can only be played once. The button used to play it is disabled after being pressed once. After the clip has been played the input fields necessary for this exercise (the text field and the input button) are activated. After each word has been entered the text field is emptied. Users are not able to see which words have already been entered. Subjects can stop recalling words and move on to the next list even before inputting a single word. There is no time limit or other form of pacing involved for the recall process. After the instructions for the first list have been shown it is no longer necessary to repeat them, so the other lists contain instructions to perform the same steps as before.

4.2.4 Exercise 2: Mitigation

Design

This part of the experiment asks participants to judge how friendly an utterance is. This is meant to test the mitigation effect of filled pauses. Subjects will be given three adjacency pairs. Of each adjacency pair there are 4 versions: computer speech with filled pauses, computer speech without filled pauses, human speech with filled pauses and human speech without filled pauses. Subjects are given two different versions of an adjacency pair and asked to compare the two replies. This is repeated four times with different combinations. The specifics of this combination are discussed in the "Materials" chapter of this exercise.

The clips contain adjacency pairs where the reply is unfavorable/dispreferred. Subjects are asked to compare the harshness of the objection/negative reply of two audio clips using one of the following 5 statements:

- "Audio 1" is much less friendly than "Audio 2"
- "Audio 1" is less friendly than "Audio 2"
- "Audio 1" is equally friendly than "Audio 2"
- "Audio 1" is friendlier than "Audio 2"
- "Audio 1" is much more friendly than "Audio 2"

The intention of this exercise is to measure the mitigation of filled pauses as well as the difference of this effect between computer generated and human speech. There is also a comparison between human and computer speech to see whether or not subjects care about the fact that it's unnatural sounding when judging friendliness.

In addition to the comparisons, users are asked a question with regards to the sincerity or the replies. The questions asked for each pair are show in Table 8 under "Additional Question". Filled pauses are generally seen as indicators for thought. It is possible listeners may interpret this as time for the speaker to come up with an excuse. The additional questions are intended to confirm or dispute this.

Materials

The adjacency pairs used in this exercise can be seen in Table 8. The four times users are asked to compare two audio fragments are used as follows:

Comparison \ Combination	I	II	III
1 – TTS vs TTS+FP	1	1	2
2 – TTS vs TTS+FP	2	3	3
3 – Human vs Human+FP	3	2	1
4 – Human+FP vs TTS+FP	1	3	2

Table 7 – distribution of adjacency pairs across the three different combinations (indicated by roman numerals) subjects can receive. The leftmost column represents the comparison number and which versions of the adjacency pair will be compared. The values to their right are numbers representing adjacency pairs in Table 8.

The three possible combinations in this table correspond to the same roman numerals shown in Table 5. Simply put, a participant that gets combination I for exercise 1 will also get combination I for exercise 2 and so on. The combinations ensure that each adjacency pair's two computer speech versions are compared an equal number of times.

The content of the sentences chosen for the exercise involved creative decisions, but their structure was consciously chosen to contain major syntactic boundaries where filled pauses could be used. These are common in the analyzed corpus. The duration was controlled in the same way as in the first exercise. The filled pauses were either in front of the entire response or in between the initial reaction and the explanation. This could affect the rating of honesty.

Adjacency pair 1:
A: Hé, wil je samen gaan lunchen? **B:** * Nee. Ik heb al geluncht.
Additional question:
Welke van de twee antwoorden klinkt het meest als een smoes?
Adjacency pair 2:
A: Ik ga wat koffie halen, wil je ook wat? **B:** * Nee.
Additional Question:
Welke van de twee antwoorden klinkt het meest als een leugen?
Adjacency pair 3:
A: Kan je me weer een lift naar huis geven? **B:** Alweer? * Vandaag kan ik niet.
Additional question:
Welke van de twee antwoorden klinkt het meest als een smoes?

Table 8 – Adjacency pairs and questions for part 2 of the experiment.

Procedure

Subjects are presented with two audio clips 4 times. Each time a user is able to listen to both audio clips as many times as he/she wishes. At any point the user can move a slider along an axis. When changing the value the label underneath the slider is updated to show what the current position of the slider represents. The choice must then be confirmed by pressing the button used to submit input. After the first 3 comparisons the additional question is asked about the perceived honesty of the replies. The input for this is given using the text field and the submit button. Instructions are given at the before and during each step as to what input fields to use and the aspect by which to compare the sound clips.

4.2.5 Exercise 3: Listener Preference

Design

This part of the experiment is another rating exercise where subjects are asked to compare two audio clips. As mentioned before, this is meant to test hypothesis 4. Preference for speech is determined by asking subjects to compare two computer speech utterances whose only difference is the presence of a filled pause in one. The comparisons the user are asked to make are:

- Which utterance sounds more natural
- Which one is easier to understand
- Which one is more pleasant to listen to

Materials

The utterances to be compared are shown in Table 9. The two audio clips only differ in that one has a filled pause and the other does not. The locations of the filled pauses within the sentences are based on the corpus analysis. In the first sentence the location is in front of a word with low accessibility. This represents the often occurring “pause to find wording” where a speaker pauses to find a single word to express the meaning he/she is thinking about. In the

second sentence the location of the filled pause is after a conjunction in a complex sentence. The corpus showed many of the filled pauses occurring at these major syntactic boundaries. The duration was manually adjusted until it was subjectively deemed to sound natural. Finally, the third sentence uses a construction found often in the corpus; it is similar to the second sentence in that it's a conjunction of two sentences. The duration of the filled pauses, as with the previous exercises is within the range of the corpus data of the appropriate boundary strength.

It is worth noting that the filled pause in the third sentence is in fact "[uh] ja" and not just the "[uh]". The alternative version of the audio contains a silent pause where the filled pause would normally be. The reason for this is to see whether the construct sounds strange to a listener when compared to "[uh] ja" and so doing influences their preference.

Dat noemen ze * mensenvrees.
Het is goed, maar * het kan beter.
Ik kwam laat thuis, en * ja, ik wilde meteen naar bed.

Table 9 – Short sentences used to judge listener preference in exercise 3. Asterisks indicate where signals are located.

Procedure

As with the second exercise of this experiment, users are able to play two audio clips as many times as they want to and answer the question based on that. Instructions are given before the exercise and during every step as to what buttons to press and what aspect of the two audio clips they should compare. Three questions are asked for each two audio clips; one for each of the comparison criteria.

4.2.6 User Selection Criteria

The only criterion for participation is fluency in the Dutch language. In this experiment as with all experiments involving language it is important to distinguish between native and non-native speakers. In this case both native and non-native speakers are allowed to participate. The distinction is made using the user information gathered at the beginning of the experiment. Persons familiar with the goal of the experiment were not asked to participate.

4.2.7 Sample Size

The goal for number of participants was set at 60. There are three versions of the experiment, referred to as 'combinations'. Due to the even yet pseudo-random distribution, 60 participants would produce 20 sets of results per version. At this point it would be possible to reasonably approximate the results of the first exercise with a Gaussian distribution. The results of exercise 2 were distributed in such a way that the "most important" types of comparisons were done twice in three combinations resulting in about 40 units if the goal was reached. The results of exercise 3 would count the same as the actual number of participants because the exercise was identical for all versions of the experiment.

It was later decided to use Fisher's exact test instead to calculate the significance of measured correlations. The actual number of participants came out to 63, of which 4 data sets were incomplete. This was compensated for when calculating averages as is explained in 4.3.1.

4.3 Experiment Results

4.3.1 Pre-processing

The data collected from the exercise is contained in a database. The database remained online and continued to operate after the experiment was concluded. In order to prevent disastrous data loss the relational database tables were output to a local text file. All calculations were done on this local file meaning that the database would be kept free from additional load and as mentioned before, the data would be kept safer. While the first calculations were being made the last few data sets were trickling in. Taking those into account as well is a simple matter of updating the local file. The text file is loaded by a separate application designed to clean and convert the data into tab-delimited tables which could then be easily loaded into a spreadsheet program.

The bulk of the database entries are comprised of data for the first (free recall) exercise. The first exercise is the only point during the experiment where user input has any degree of unpredictability. As a result the data contains errors that needed to be corrected/cleaned before the actual statistics could be taken. The errors that needed correcting were:

- Multiple words on one line
- Multiple entries of the same word
- Misspelled words
- Misaligned numerical order of recalls

The first step in cleaning the data was splitting entries where strings contained commas or spaces. Certain exceptions needed to be left un-split, because they contained a single entry despite the presence of a space. These exceptions were hard coded as a pre-condition in order to leave them unchanged.

The second step was correcting misspelled words. This is a crucial part of the data cleaning process, because this is also the point where words that were *clearly* misunderstood were replaced with the correct word. This could be considered tampering, but because words in the lists are distinct enough and the corrections are equal for all versions of a list the chance of false positives is kept to a minimum. A false positive, in this case is an entry where the subject recalls a word that in reality he/she did not. Correcting misspelled words has the same rate across the different versions compensating them each equally. For those words of which it was unclear what word they were *supposed* to be, if any at all, were removed and not counted.

The problem of multiple entries of the same word is somewhat similar to the problem of recalled words that did not appear in the list in the sense that both should not be considered when counting word frequencies that lead to the recall-rate numbers. However, when looking at the data it becomes immediately apparent that these words are not creations of the subjects but rather words they heard in previous lists, so called *'misplaced'* words. In some cases these words were not originally recalled during the recall period designated for that list. As is discussed in chapter 4.3.3 the misplaced recalls are not very common and, because many of them are duplicates, only affect the results by a negligible amount. Duplicate words are quite common. Although the pure recall rate should not count the same word more than once, a check is done for the existence of a relationship between duplicate recalls and filled pauses.

Exercises 2 and 3 of the experiment had multiple-choice style questioning which results in predictable data format and contents. The tables for these exercises were loaded without any modifications, because they needed no pre-processing.

Although 63 persons participated in the experiment resulting in 21 iterations of each audio combination (see Materials section for exercises 1 and 2 in chapters 4.2.3 and 4.2.4) the data for a few participants is incomplete. The cause for this was most likely the disruption of communication or closing of the experiment application while it was still transmitting data. The number 21 can not be used reliably for calculating averages for each combination, so every time an average needed to be calculated a separate count of the number of results was done. To summarize, these counts were made for each of the following: each version of each list in exercise 1, each question type of each adjacency pair in exercise 2 and every individual question in exercise 3.

4.3.2 Subject Demographics

Participants were asked for personal information in order to determine possible links between effects and effectiveness of filled pauses on different demographics. The resulting composition of the 63 participants is displayed in the following table.

Version 1 (Group A)					
Age	**Under 18** 1	**18-30** 13	**31-40** 2	**41-60** 3	**60+** 2
Gender	**Man** 14	**Woman** 7			
Language	**Native** 14	**2nd lang.** 7			
Experience	**None** 4	**Little** 14	**Regular** 3		
Version 2 (Group B)					
Age	**Under 18** 0	**18-30** 16	**31-40** 2	**41-60** 2	**60+** 1
Gender	**Man** 9	**Woman** 12			
Language	**Native** 16	**2nd lang.** 5			
Experience	**None** 7	**Little** 10	**Regular** 4		
Version 3 (Group C)					
Age	**Under 18** 0	**18-30** 15	**31-40** 1	**41-60** 4	**60+** 1
Gender	**Man** 14	**Woman** 7			
Language	**Native** 15	**2nd lang.** 6			
Experience	**None** 6	**Little** 11	**Regular** 4		

Table 10 – composition of participants of the filled pause experiment grouped by the version of the experiment they completed.

Table 10 shows the composition of the participants for each of the three versions of the experiment. Each group consists of 21 participants due to the even distribution of the versions. All age groups other than the 18-30 demographic had low numbers. For this reason when presenting the results of each exercise in the following chapters, the participants are split into two age groups: 30- and 30+.

4.3.3 Exercise 1: Word Highlighting

Recall Rate Increase for Signaled Words

After pre-processing, the data for exercise 1 contains only words that appeared in the list, correctly spelled and without duplicates. The simplest representation of the recall rate is a table with the count for each word in each list; such a table is displayed in Appendix C. For ease of reading the words preceded by filled pauses are displayed in bold text. The columns represent each version of the list and give the first impression of differences between them. As

mentioned in the previous chapter, not each version of each list was completed an equal number of times. The number of times is displayed at the bottom of each column. This value ranges from n=19 to n=21. When the number of times a word is recalled for each version of a list is divided by the total number of times that list was results in a chart like the one in Figure 12.

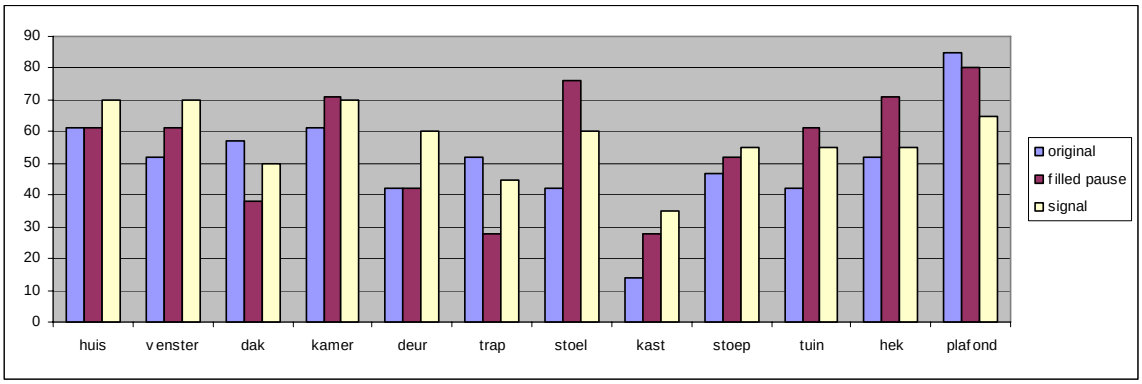


Figure 12 – a chart depicting proportional frequency of word recalls for list 2. The series are in order from first to last: unmodified, filled pauses and acoustic signal.

As a reminder, the hypothesis this exercise attempts to prove is: *A filled pause in front of a word causes heightened attention which improves the chance that a listener will be able to recall it.* The recall rate of each word in list 2 is shown in Figure 12. This table shows the most dramatic increase in recall rate for any single signaled word in all of exercise 1. This word ("stoel") had an average recall rate of 42% over 21 iterations of the unmodified list, which is clearly dwarfed by the 76% average recall rate over the same number of lists with filled pauses. The other word in this list that was preceded by a filled pause ("kamer") has a much smaller increase over the unmodified version. For both words the recall rate when signaled by an acoustic signal lays between the rates for the unmodified and filled pause versions.

These increases in recall rates suggest an increase in the likelihood of a word being recalled when preceded by a filled pause than when it is not. Additionally, the effect of the acoustic signal is similar but less effective than that of the filled pause. These facts are in line with the expectations, showing the unique and positive effect of filled pauses. However, these findings are unique to this list alone. Unfortunately the trend shown in list 2 is not universal across all lists.

Each of the 6 lists contains two possibly signaled words, adding up to a total of 12. Consider the alternative hypothesis: filled pauses have no effect or a detrimental effect on the recall rate of words following it, i.e. the variables filled pauses and recall rate are independent. If this were true then, all other things being equal, it is probable that out of the 12 cases of signaled words half or more than half would exhibit an equal or greater recall rate for words not preceded by a filled pause. This is however not the case. 7 out of 12 cases show a greater recall rate for words signaled by a filled pause. Out of the 5 that did not, 2 had equal recall rates and 3 saw a decrease in recall rates when signaled by a filled pause. The significance of all this can be measured by calculating the p-value. In short, given that there is no relationship between the variables, the p-value is the chance of measuring a difference equal or greater than the one measured in the experiment. The p-value is calculated using a 2x2 matrix and Fisher's exact test.

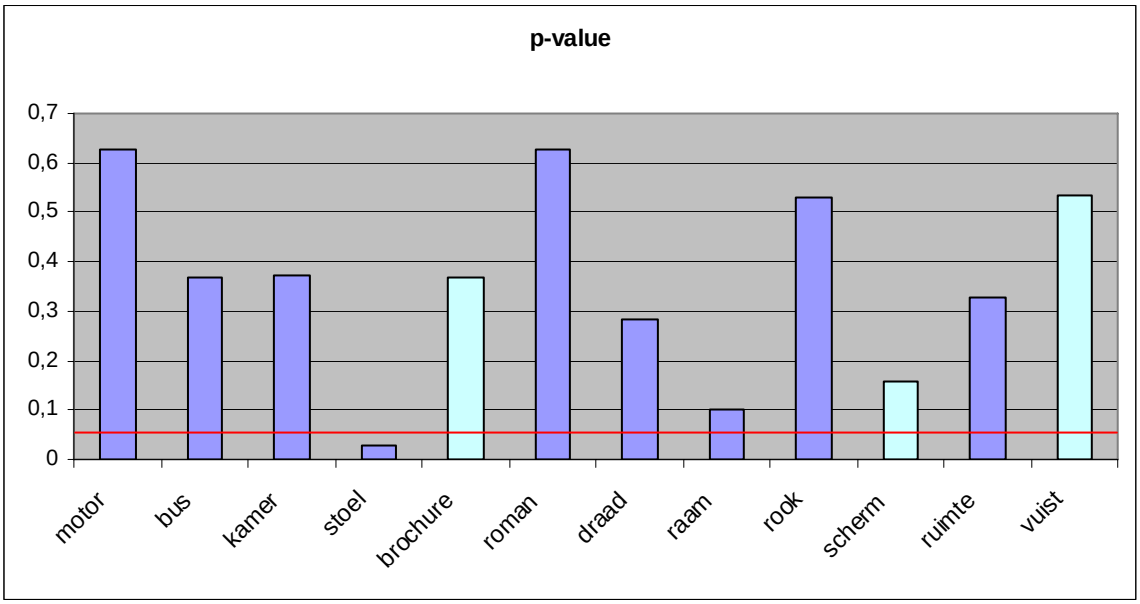


Figure 13 – p-values of the measured relationship between filled pauses and recall rate displayed for each signaled word. The red line shows the 5% maximum for significance. The light blue columns show indicate the three cases where the correlation was negative.

As can be seen in Figure 13, only one of the measured correlations is of statistical significance when using the standard 5% as a maximum. This increase was that of the word "stoel" in list number 2 (see Figure 12). Most of the other p-values are quite high and are too likely to have occurred by chance to be the basis of any conclusions with regards to the effects of filled pauses. As mentioned before, three of the 12 signaled words showed a decrease in recall rates when preceded by a filled pause. These three cases are shown in Figure 13 with light blue bars. The p-values measured for these three correlations have the same meaning as the others, but for a negative correlation between the two variables.

Consider the factors that could have influenced the recall rate of the word "stoel" other than the filled pause. The group of participants that received the filled pause version of list 2 (now referred to as group A) also received the unmodified version of list 1. The group that received the unmodified version of list 2 (now referred to as group C) also received the acoustic signal version of list 1. This means that the latter group was in fact more prepared for any signaling than the former. Assuming that being surprised by the signal would have a detrimental effect on the recall rate it would be expected to see a higher recall rate in the group C. However, the opposite appears to be true. Another outside factor could be the composition of the two groups. On closer inspection, there are no parameters in the user information that varies substantially across group A and C. The user information for all three groups is listed in Table 10. Ruling out these factors is not enough to conclude that the increased recall rate of the word "stoel" was due to the filled pause, because there is still a 5% chance of finding a statistically significant correlation by chance. Given the number of signaled words (12) a single statistically significant result is likely.

Variation among Genders

The gender distribution was slightly uneven with 37 males and 26 females participating in the experiment. The males and females are distributed among the different versions as indicated in Table 10. It shows that the smallest group when separated by gender counts 7 females (Groups A and C). In both cases that represents 1/3rd of the total participants in that group. Because of the smaller number of participants per group when separating by gender, the results will show more variability and will require a stronger correlation to be considered statistically significant.

The overall serial position curves for all groups don't deviate very much when they're represented visually. In some specific cases, however, they do. Certain specific words seem to have a distinct recall rate across the genders. As an example, consider the word "bureau" in list 5. Looking at the unmodified version of this list will exclude the effects of any signals. In the unmodified version of list 5 the male participants recalled it 1 out of 9 times while the female participants recalled it 7 out of 11 times. This difference is statistically significant with a p-value of 2.4%, which, as before, was calculated using Fisher's exact test. Similarly, the words "Mercedes" (list 1) and "pion" (list 6) have a statistically significant correlation between gender and recall rate. This could be interpreted as evidence that men and women recall certain words at a different rate, or alternatively, it could suggest that men and women have a different strategy they use to remember items from a list. However, if the chance of finding a statistically significant correlation by accident is set at 5%, then out of the 72 words in the experiment the expected number of erroneous significant results is 3.6. The 5% chance of a false positive applies to each word individually, but is equal across all of them because the same significance limit is used in all of the experiment.

The changes in recall rates across different versions of the lists also contain some variation between males and females. In list 1 the filled pause before the word "motor" apparently had enough of an impact on the male participants to ensure it was recalled 0 out of 9 times. As a result the change in recall rate for the males was -35%, while the females experienced an increase in recall rate of 30%. The recall rate changes vary this much because of an underlying cause. The underlying cause being the difference in recall rates for the word "motor" when preceded by a filled pause for males and females. The p-value for the measured relationship between these two variables is 0.6%. In the same list the other signaled word "bus" has a relatively big difference between recall rate changes. However, this time the male participants have the recall rate increase and the females have the recall rate decrease. The two signaled words in list 1 have a similar but reversed change in recall rate across the genders. They are also the largest changes for all words in the list. In the other lists changes in recall rates vary across the board. Because of the spread there is no evidence to suggest that men and women interpret filled pauses differently. As for acoustic signals, the same conclusion applies.

As for overall performance the females seem to consistently outperform the men. By counting the total number of recalled words for each version of a list and dividing it by the number of males or females and then averaging the results of all 6 lists you get the following results:

	Males	Females
Unmodified	5,41	5,75
Filled pause	5,36	5,75
Acoustic Signal	5,09	5,84

Table 11 – overall performance of males and females for exercise 1. Values represent the average number of recalled words per list. The table shows that females outperformed males in all versions of the list.

Variation across Language Dominance

Of all the participants 18 said Dutch was their second language, the other 45 specified it as being their native language. When subdividing the groups by language dominance recall rates vary less than when separating by gender with the exception of the category of lists accompanied by acoustic signals. In those cases non-native speakers seem to score better than native speakers.

The word "Mercedes" is an internationally renowned brand name. As expected this word's recall rate is very similar with native and non-native speakers. Some words have a significantly higher recall rate in one group. One such word is "stoep" in list 2. Non-native speakers recalled "stoep" 6 out of 6 times while native speakers recalled it 4 out of 15 times. There are more of these, but the total number is consistent with the 5% error rate.

The differences in recall rates are spread out as are the differences in recall rate changes. The lack of consistency or tendency suggests that there is no difference in how native and non-native Dutch speakers interpret filled pauses. Nor is there evidence that one group outperforms the other in this free recall exercise.

Variation among Age Groups and TTS Experience

The other two parameters by which the participants can be split produce very small groups. As can be seen in Table 10 the best represented age group is 18-30. As for experience with computer generated speech, the groups are not as badly represented as with the age parameter, but still quite small, ranging from 3 to 14 people.

The recall rate of an individual word will be skewed because of the small numbers, so instead the average recall rate of all signaled words is shown in the following table.

Category \ Age	up to 30 (45 participants)	31 and over (18 participants)
Unmodified	33%	30%
Filled Pause	37%	44%
Acoustic signal	31%	39%

Table 12 – average recall rate of the 12 words possibly accompanied by signals. Values are grouped into two age categories.

The most notable difference in the average recall rates is the high recall rate for the higher age group when signaled words were signaled by a filled pause. The acoustic signal also seems to increase the recall rate for the second age group. The average recall rate of the unmodified version of the signaled words don't differ very much, but it does support the generally accepted assertion that memory performance is negatively correlated with age.

Category \ Experience	None (17 participants)	Some (35 participants)	Regular (11 participants)
Unmodified	35%	34%	31%
Filled Pause	29%	44%	43%
Acoustic Signal	29%	33%	44%

Table 13 – average recall rate of 12 signaled words grouped by experience with TTS.

Table 13 shows the average recall rate of signaled words grouped by their experience with computer generated speech as specified by them before performing the experiment. The unmodified versions of the words seem to be recalled at an equal rate by all 3 groups. The small differences are somewhat surprising considering the relatively small number of participants in the two outside categories ("none" and "regular"). The most notable difference is the general decrease in recall rate for the first group when the words are signaled compared to the general increase in the group with the 'regular' experience with computer speech. This increase in recall rate for the filled pause signals can be explained by a stronger ability to distinguish a filled pause from the rest of the word. The same argument can be applied to the second group. More experience with filled pauses could also have caused a lack of surprise when hearing the filled pause. However, lack of surprise alone may not explain an increase recall rate. The difference in average recall rate for the words accompanied by an acoustic signal is quite unexpected. This could have an underlying cause such as the most experienced users listening for anything unexpected, while the others were being more concerned about the overall performance.

Subdividing by age and experience both produce small groups whose composition could be uneven. The average recall rates are given as one of the more reliable measures which would average out any skewing caused by other factors such as the composition of the small groups. The results are not dependable but are an indication of possible tendencies. Knowledge of such tendencies facilitates the correct interpretation of other results and design of future studies.

Recall Rate of Non-Signaled Words

The average recall rates for all words reveal no tendency in either direction. There is no evidence to suggest that signaling affects the overall performance in a free recall exercise negatively nor positively. The difference between the average recall rates of unmodified and filled pause versions of each list are shown in Figure 14. There is no clear consistency. However, when this same comparison is made between the unmodified version and acoustic signal versions of each list (see Figure 15), there is an apparent tendency. The acoustic signals seem to negatively affect the overall performance of the subjects. This effect averages out to about -2.7%

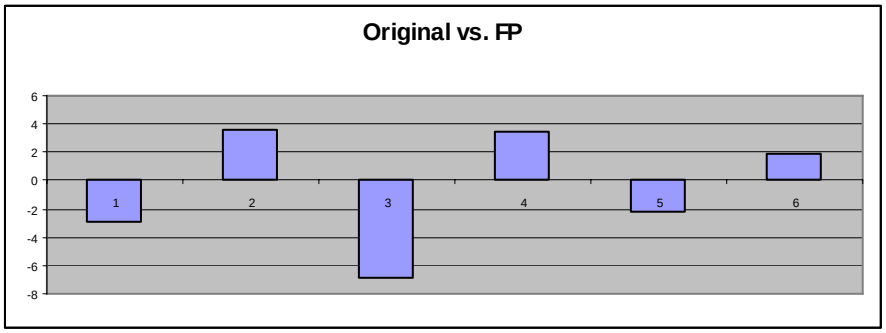


Figure 14 – The bars represent the difference between average recall rates for all words in each of the 6 lists. The differences are not consistent.

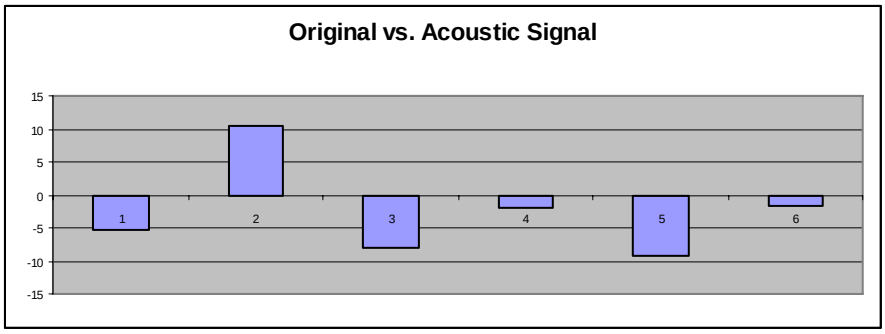


Figure 15 – the difference between performance of the original and acoustic signal version of the free recall exercise.

One of the conditions for the placement of signaled words was to separate them with at least one other word to prevent the heightened attention to overlap and make measurement less precise. Temporization of the presentation rate could affect the amount of rehearsal done by the subject which would affect the recall rate of words preceding an element that causes such temporization. The recall rate of the words preceding the first filled pause in each list are averaged and compared. The results in both cases show a tendency that contradicts the theory of additional rehearsal time. Both types of signaling seem to negatively affect the recall rate of words preceding the first signal, filled pauses more so than acoustic signals. Filled pauses on average reduce the recall rate by 4% while the acoustic signal does the same by a negligible 1.5%. Hearing a filled pause apparently affects the focus but also the 'working memory'. It could also disrupt the subject's strategy for remembering words, which would result in a decrease in the likelihood of a recall later on.

The effect on words following the second signaled word could be considered an extension of the effects on signaled words. It was assumed that the heightened attention this experiment is meant to confirm has a short duration, but the following results suggest this is not the case. Words that follow beyond the signaled words are grouped and their recall rates averaged. The result of comparing the unmodified and filled pause versions of each list is shown in Figure 16. The chart shows a tendency for an increase in recall rate for words following a word signaled by a filled pause. This tendency averages out to 5.4%, while on the other hand the acoustic signal shows an opposite tendency, decreasing the recall rate by an average of 3%.

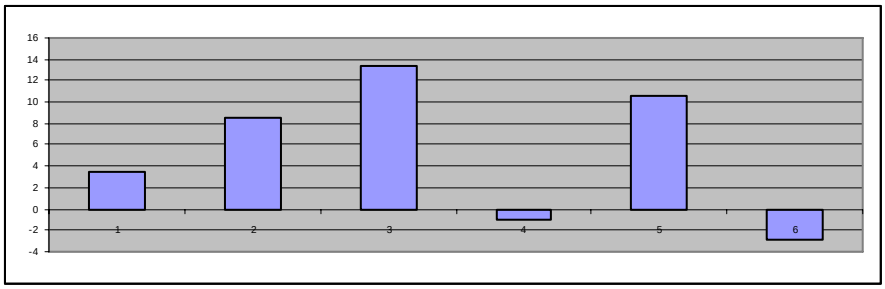


Figure 16 – The difference between average recall rates for all words following a signaled word for all 6 lists. The differences indicate an increase in recall rates.

This effect is consistent with the exception of lists 4 and 6, whose decrease in recall rate is relatively low. When the recall rate of the last signaled word is added to the sum of the words

following it the consistency and average increase is even better as can be seen in Figure 17. This increase of recall rate does may come at the cost of the recall rate of the other words in the list. However, the recall rates of the words preceding the second filled pause only decrease in 4 of the 6 lists. The decrease in performance for the group of words preceding the second filled pause is not as pronounced as the increase for those following it, but it on average there does seem to be a decrease.

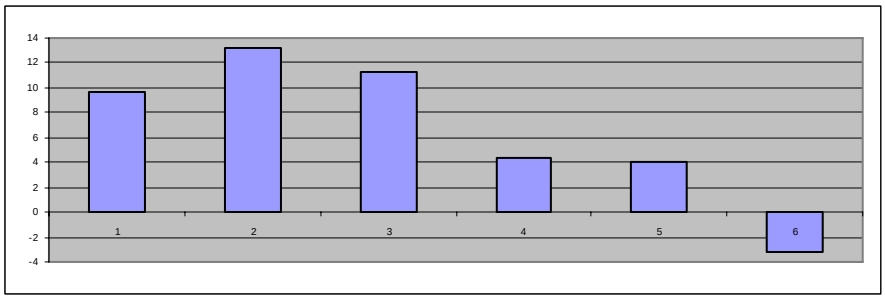


Figure 17 – The difference between average recall rates for all words following the last filled pause. Each bar represents the increase for one of the 6 lists. The chart shows a tendency of the recall rates to improve.

The versions of the lists containing the acoustic signals did not outperform those with filled pauses as mentioned before. The recall rate increases of the same group of words that resulted in Figure 17 produces a much different picture for acoustic signals. The equivalent of that chart is shown in Figure 18. The chart for acoustic signals shows no consistency for either increase or decrease of recall rates.

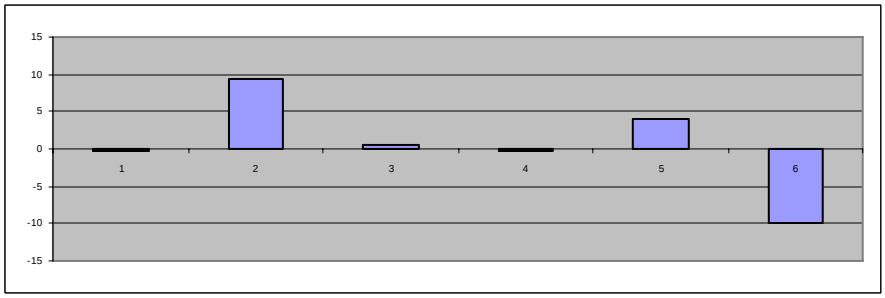


Figure 18 – average recall rate increase for all words following the second *acoustic signal* in all 6 lists. The chart shows no trend for either increase or decrease in contrast to Figure 17.

Misplaced Recalls

During the free recall there were several cases of subjects recalling a word that was present in a previous list. This phenomenon occurred 30 times, of which 10 saw those words being recalled for the first time by that person. Several of the misplaced words are the same word. For example, the word “venster” was recalled six times outside the proper recall moment. Of these six misplaced recalls four happened during the recall period of the 4th list. That list contains the word “ventiel” which could possibly be the cause of the misplaced recall. Similarly, the word “patroon” was misplaced five times in lists 5 and 6 possibly due to the word “potlood” in list 5.

Both examples have cases where the word was recalled for the first time. This phenomenon is interesting in that similar words could possibly be interpreted as hints for words subjects previously heard but could not recall on their own. However, this phenomenon is not common enough to drastically change the overall performance of subjects. Only one of the signaled words appears in the list of misplaced recalls and it was also recalled at the proper time.

Order of recall

The order in which words were recalled is used to deduce a measure of a word's accessibility. Every recall has a value that represents its place in the series of recalls it is part of. This weighted mean of this position is calculated for each of the signaled words for each modification of the list. The result can be seen in Table 14. The 'level of accessibility' does not consistently increase or decrease under the influence of a filled pause or an acoustic signal.

Word	Unmodified	Filled Pause	Acoustic Signal
motor	5	4	4
bus	4,2	4	3,7
kamer	4	4,6	3
stoel	4	4	4,75
brochure	5	4,8	4,2
roman	4	4	5
draad	3	6	3,75
raam	5	4	3,7
rook	4	5,75	2
scherm	6	3	3
ruimte	4	4,25	4
vuist	2,75	5,75	4

Table 14 – weighted mean of the position of a recall in its series calculated per word per version of the lists.

Conclusion

The recall rate of signaled words increases on average, but only does so 7 out of 12 times. Additionally, only one signaled word experienced a statistically significant increase which is not entirely unexpected considering the number of signaled words and a likelihood of 5% of a false positive. The results for the signaled words are inconclusive. When considering signaled words, there is not enough evidence to support hypothesis 1. However, the tendency to increase recall rate suggests that such evidence may indeed exist, in which case the effects of the filled pause are too subtle to be measured with the performed experiment.

The effects of the filled pauses were assumed to be limited to only the following word. However, the increase in recall rates of all words following the last filled pause were put in perspective. The recall rates of these words increased in 5 of the 6 lists, of which the exception had minimal decrease. The increase in recall rate ranged from -3% to +13% and averaged at 6.5% (see Figure 17). In contrast to the signaled words, when looking at all words following a filled pause the results of exercise 1 support hypothesis 1.

The free recall exercise also contained lists that had words signaled by an acoustic signal. The 3rd hypothesis was formulated to make a specific comparison between a non-verbal acoustic signal and a filled pause. The acoustic signal, though chosen so as not to interfere with the process of memorizing the list items, is out of place in the list and can produce adverse effects.

This concern is can be dismissed by comparing the recall rate increases of all words following the second signal of each list. The acoustic signal does not have any consistent tendency to either increase or decrease the recall rates. The experiment's results suggest that the positive effect of filled pauses on the recall rates is indeed stronger than that of the acoustic signal. The experiment supports hypothesis 3.

4.3.4 Exercise 2: Mitigation

This part of the experiment is a survey intended to measure the effects of perceived friendliness of replies in adjacency pairs.

Change in Perceived Friendliness

The different combinations given to participants resulted in an even distribution of all adjacency pairs for each comparison. However, there were two comparisons made between synthetic speech with and without a filled pause in each combination. The results of exercise 2 are included in Appendix D, where the last column of the first table shows how many iterations of that comparison were made for each adjacency pair. The comparisons, that had as only difference the presence filled pauses, indicate the direct effect it has on the perceived friendliness. This comparison was made for both synthetic speech and human speech. The results are shown in Figure 19. The series shown in the charts' legends represent from top to bottom: much less friendly, less friendly, equally friendly, friendlier and much friendlier. These ratings compare the first type of reply against the second as shown in the title. For example, the top chart in Figure 19 shows that almost 70% of all participants rated the filled pause version of the reply in adjacency pair 2 as being "friendlier" than the unmodified version.

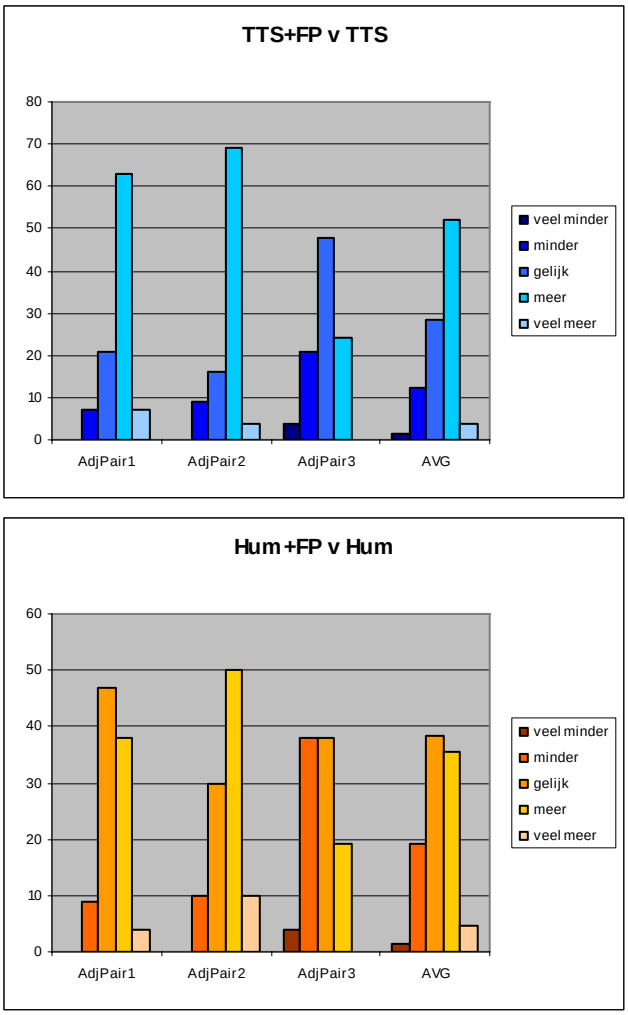


Figure 19 – charts with the frequency of each rating category grouped by adjacency pair. The top table shows the ratings comparing synthetic speech with filled pauses against synthetic speech without. The bottom table does the same for human speech.

The results show that most subjects preferred the speech with the filled pause in two of the three synthetic speech replies. In the 3rd adjacency pair the majority had no preference. The differences between the ratings for human speech indicate that either the filled pause in synthetic speech has more effect in mitigating the rejection of the replies than in human speech or that human speech without a filled pause is friendlier than the synthetic speech without. In either case the improvement in friendliness of the reply when adding a filled pause is higher in synthetic speech. The charts show the distribution of the ratings rather than a weighted mean. If the latter were used there would be only one case where adding a filled pause decreased the friendliness of the reply. That case being the human speech version of adjacency pair number 3.

The tempo and pitch of the synthetic speech was based on its human speech counterparts to make them as similar as possible given the tools and methods. The similarity is paramount to the significance of the comparison's results. The approximation of the human speech can be considered successful given that the comparison of the synthetic speech and human speech

versions of the adjacency pairs shows no definitive bias for human or synthetic speech (see the Human+FP vs TTS+FP section of the first table in Appendix D and Figure 20).

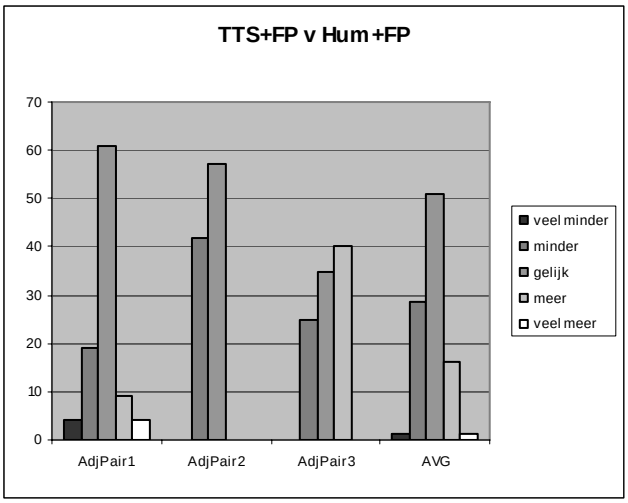


Figure 20 – results of the comparison between human and synthetic speech with identical contents. There is no consistency in the results and on average most people find the two replies equally friendly.

Change in Perceived Honesty

The additional questions asked about the adjacency pairs are meant to give and indication of how the filled pauses affect the perceived honesty of the second speaker. They were posed because it was thought that the effect on the perceived honesty could be part of the explanation for the ratings presented in Figure 19. This could be the case. The honesty ratings of the replies shows subjects find the filled pause versions sound more honest. The results are shown in Figure 21.

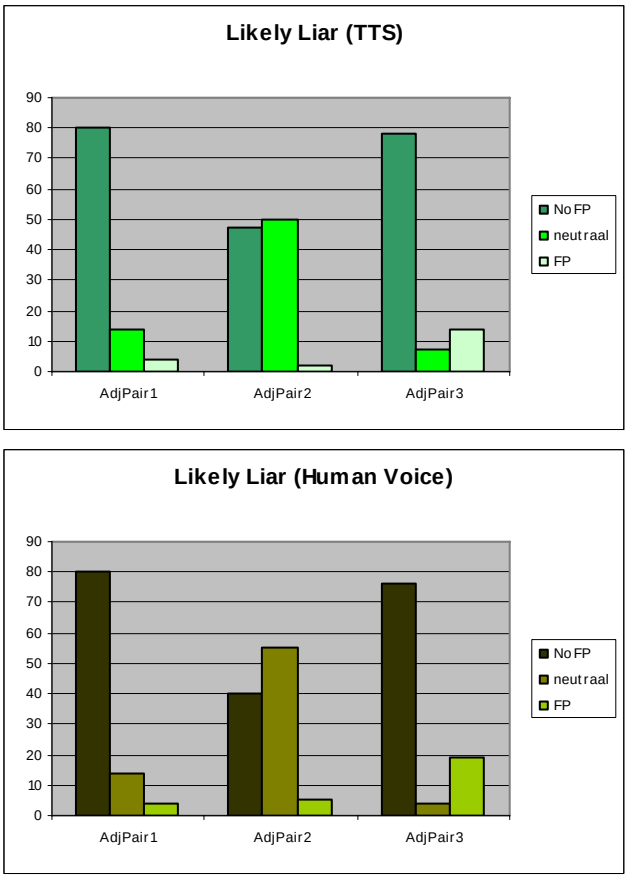


Figure 21 – rating of honesty of reply in adjacency pairs. Subjects tended to rate the non-filled pause (unmodified) version of the replies as sounding the most like a lie, indirectly judging the filled pause versions as the most honest.

There is a high degree of consistency between the human speech and synthetic speech comparison's. The distribution of the choices is similar in both charts. Filled pauses positively affect the perceived honesty in both natural and synthetic speech. The second adjacency pair is the only case where a relatively large group of participants

Variation among Genders

The first type of comparison of adjacency pair replies pits two synthetic speech replies against each other, one of which contains a filled pause. The distribution of ratings for male and female participants is very similar resulting in charts very much like the top one in Figure 19.

The comparison of human speech replies of which one has a filled pause does not exhibit the striking similarities of the first comparison. This second rating exercise shows females being consistently harsher with the reply containing a filled pause, moving their rating distribution to the left of the spectrum. This difference can be seen in Figure 22.

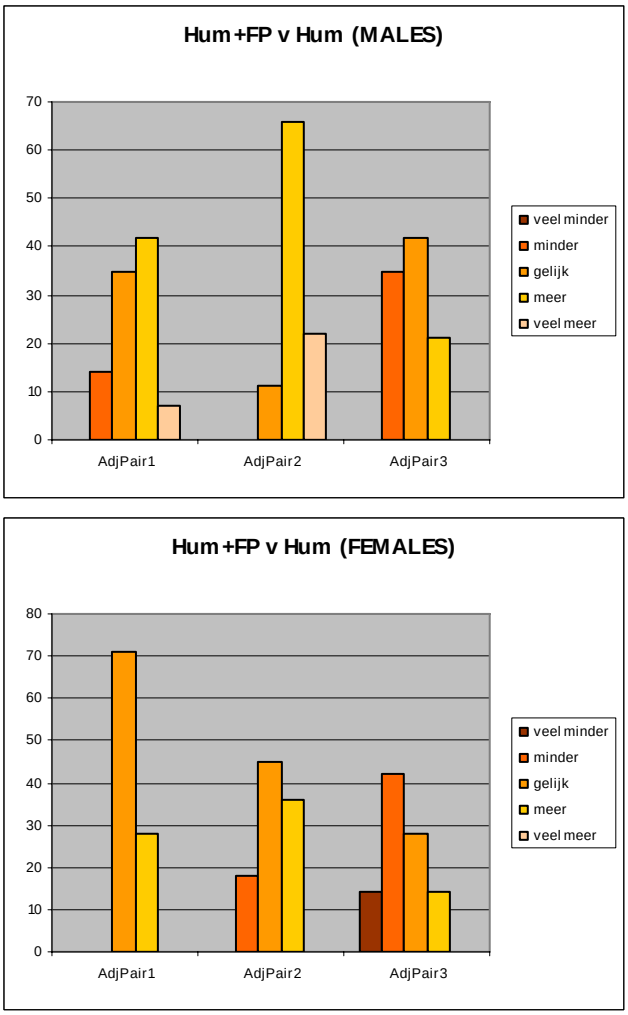


Figure 22 – comparison of human speech replies for male (top) and female (bottom) participants. Females consistently rate the filled pause reply less favorably than males.

In the last comparison of the second exercise the male and female participants don't consistently differ in their preference for either human or synthetic speech versions of the replies.

The rating of honesty of the replies does not show consistent variation across the genders. There is one case where there was an unlikely difference between the ratings of males and females. The rating of the most likely liar for the human speech reply of adjacency pair 2 had a much higher number of neutral ratings among male subjects. However, the significance of this difference is very low because of the small number of participants participating in that specific comparison (9 males and 11 females).

Variation across Language Dominance

The group of participants with Dutch as their second language is small, which means the only comparisons that can be made that hold any value are the differences between the rating distribution for the "TTS+FP vs. TTS" comparison and the "TTS honesty". The difference across the groups is minimal in both types of ratings.

Variation among Age Groups and TTS Experience

The group of participants over the age of 30 is small. The same logic applies to the restricted view of the results as with the separation by language dominance. The older group consistently rates the synthetic speech version of the replies less favorably than the younger group. The ratings distributions for "TTS honesty" are roughly the same for both age groups.

The outside groups for experience with computer generated speech (those who selected "none" and "regular" is very small. Consequently, comparing any of the rating distributions is not very practical.

Conclusion

The adjacency pairs in the second exercise varied only in their usage of filled pauses. Although this does constitute additional linguistic effort, the mitigation of negative effects in actual human speech is usually achieved through more complex constructions and other forms of hesitation. Despite of this discrepancy, participants did generally prefer the replies containing filled pauses with regards to friendliness. More precisely, two of the three adjacency pairs exhibited this property and the third one showed little to no preference either way. The results of the comparison for human speech show the effect of the filled pauses being weaker in all three adjacency pairs. The 2nd hypothesis was formulated under the assumption of computer generated speech and it is under this condition that the results support the hypothesis. The results also suggest filled pauses have a positive effect on the perceived honesty of the speaker.

4.3.5 Exercise 3: Listener Preference

Comparing FP and non-FP

Subjects were asked to compare the utterances in exercise 3 based on naturalness, understandability and which one they would rather listen to (from now on referred to as "overall preference"). The results are included in a table in Appendix E and a visual representation is shown in Figure 23. The hypothesis for this exercise expects to see no preference between the filled pause and non-filled pause versions of the sentences. The results of the first sentence support the hypothesis. Under the hypothesis the chance of a user picking the non-filled pause version is equal to the chance of picking the filled pause version. Taking the neutral answers out of the picture makes it possible to test this hypothesis.

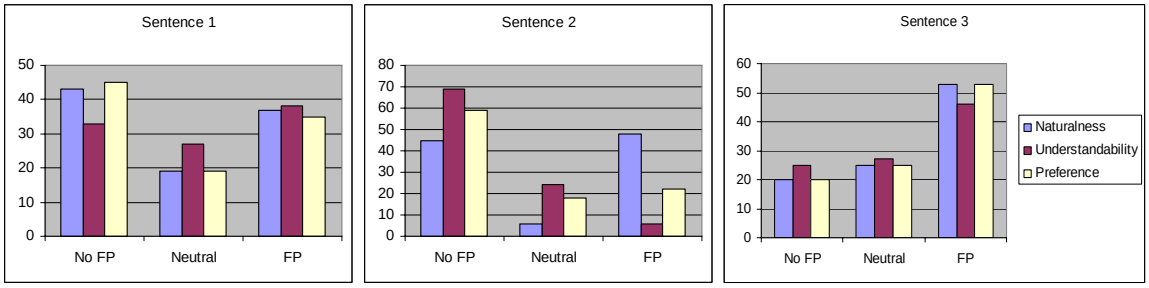


Figure 23 – charts containing the choices of subjects for each sentence in the categories "naturalness", "understandability" and "overall preference". The legend in the third chart shows which color bar represents which category.

The null hypothesis says the chance of preferring the filled pause version equals ½. A p-value below 5% suggests there is enough evidence to reject the null hypothesis, which means there is

likely to be a correlation between the rating and the presence of the filled pause. The number of times the filled pause version of the sentence was chosen is a binomially distributed variable labeled X. The p-value is calculated using the normal approximation of variable X with continuity correction as shown in the following formula.

$$P(X \leq k) = P(X < k+1) \approx P(Y \leq k+\tfrac{1}{2}) = P\left(Z \leq \frac{k+\frac{1}{2}-np}{\sqrt{np(1-p)}}\right)$$

The resulting p-values are shown in Table 15. Only one of the measured values for variable X shows statistically significant difference between the choice for filled pause and non filled pause versions of the sentence when using the 5% minimum. This value corresponds to sentence 2 in the category of understandability. Three other p-values are low and would also qualify for statistical significance if a 10% minimum was used. These values appear in each of the three categories. The measured preferences for sentence 1 show no evidence to reject the null hypothesis. Although the overall preference for sentences 2 and 3 both suggest the chance $p \bullet ½$ they *contradict* each other as to whether or not $p > ½$ or $p < ½$. This suggests the effect of a filled pause on the overall preference is situational at best. The lack of effect on preference in sentence 1 supports the null hypothesis and the opposite effects in sentences 2 and 3 lack consistency. Lack of evidence to reject the null hypothesis is *not* the same as evidence to support it. The relatively low p-values of the measured occurrences can be interpreted as signs that the rating of the different sentences does not result in a measurement of the same phenomenon; i.e. filled pauses affect the overall preference *differently* for each sentence.

As for the other categories, the preference in the "naturalness" category is not significantly lower for the filled pause version of any of the three sentences. In case of the 2nd sentence the preference regarding "naturalness" is balanced despite the other categories being unbalanced (see center chart in Figure 23). The filled pause version of sentence 3 is preferred more often in the "naturalness" category. As for the overall preference, the effect of the filled pauses on the preferences is not consistent across all sentences. In the "understandability" category the preferences between filled pause and non filled pause versions appear correlated with the overall preference.

Category \ Sentence	1	2	3
Naturalness	51%	53%	8%
Understandability	44%	2%	15%
Preference	40%	8%	7%

Table 15 – table of p-values for choices in exercise 3. Values represent the likelihood of the getting the rating results as measured in exercise 3 under the null hypothesis. The smaller p-values suggest the null hypothesis is wrong.

Variation among Genders

The overall preference for sentence 1 has opposite tendencies for males and females. Females preferred the filled pause version more frequently than males. Using Fisher's exact test we can find a p-value. The table used is shown in Table 16. The p-value represents the likelihood of measuring this or a stronger correlation given that there is no correlation. A small value suggests we can reject the assumption that there is no correlation. The resulting p-value equals 3%. This is low enough to assume there is a correlation between the variables.

Version \ Gender	Male	Female
Filled Pause	9	13
Unmodified	20	8

Table 16 – Table used to confirm correlation between gender and preference of the two versions of sentence 1.

The same calculation when done for the understandability results in a p-value of 8%. This is in line with the suspected correlation between understandability and overall preference. The same calculation when done for the overall preference in the other two sentences produces high p-values. This can be explained by the correlation found being a false positive or unique to that sentence/circumstance. Sentences 2 and 3 have very similar results for male and female participants.

Variation across Language Dominance

There are no notable correlations to be found between language dominance and the distribution of the choices with the exception of naturalness of sentence 1. However, the p-value from Fisher's exact test equals 7%. This is not statistically significant nor is it consistent with the rest of the results. The content of sentence 1 doesn't reveal any particular reason why native Dutch speakers would prefer the filled pause version more often than the other group.

Variation among Age Groups and TTS Experience

For the age groups "30 and younger" and "31 and older" there is only one statistically significant correlation to be found. It belongs to naturalness of sentence 2. Looking at the sentence there is no obvious reason to explain why the older group would prefer the non filled pause version more often with regards to naturalness than the younger group.

When separated by experience with computer generated speech the groups exhibit only one anomaly. As with the grouping by age, it's the naturalness of sentence 2 where one of the groups differentiates itself. The smaller outside groups ("No experience" and "Very experienced") show a tendency to prefer the filled pause version in this category while the middle group shows an opposite tendency. This correlation has no explanation to be found in the sentence itself or the nature of the grouping and is likely a false positive occurring purely by chance.

Conclusion

The preference measured in exercise 3 did not result in any consistent results. The effects of the filled pause were very apparent in the overall preference for two of the three utterances, but the effects were opposite. Although averaging the results of the three sentences would support hypothesis 4, the occurrence of the two extremes in the same exercise suggests there is a more complex underlying process. The conclusion that can be made given these results is that the context of an utterance determines if listener would prefer hearing a filled pause over not hearing one in the utterance.

5 Conclusions

5.1 Properties of Filled Pauses in Natural Language

The first part of this paper focused on the corpus analysis which produced results regarding the duration of filled pauses, their frequency and a list of causes for their occurrence. The results regarding their duration can be summarized in the following points.

1. The average as well as the standard deviation of the duration of filled pauses at clause boundaries (major syntactic boundaries) is very similar across the 5 speakers.
2. Conversely, the distribution of the duration for filled pauses at minor syntactic boundaries is very variable across the speakers.
3. The standard deviation of the duration of this last group of filled pauses appears to be correlated to the frequency at which a speaker uses filled pauses.

The frequency of filled pauses is very inconsistent across the speakers in the corpus as is to be expected. However, the distribution of filled pauses among major and minor syntactic boundaries is similar. As mentioned in the results regarding duration, a relationship was found between a speaker's frequency of filled pauses and the variability of the duration of filled pauses at minor syntactic breaks. To summarize:

1. Speakers use filled pauses at highly varying rates.
2. The *proportion* of the filled pauses appearing at minor syntactic breaks is consistent among speakers. Roughly 3 out of 5 filled pauses appear at minor syntactic breaks.

Subdividing the filled pauses into causal categories showed that the vast majority of them occurred to select the wording appropriate to the meaning the speaker wished to express. The categorization is admittedly subjective.

5.2 Effects of Filled Pauses in Synthetic Speech

The second part of this paper presented an experiment which combined the data about the duration and form of filled pauses gathered from the corpus analysis and tested the effects of filled pauses in certain functional roles as gathered from other studies. The experiment tested several hypotheses to verify the word highlighting and mitigation functions of filled pauses in computer generated speech. The findings of the first exercise of the experiment are:

1. The free recall exercise did not show whether or not a single word immediately following a filled pause benefitted from its presence in terms of recall rate.
2. Filled pauses *do* affect the overall recall rates of *all* words that follow the second signal in each list. This effect is an *increase* in average recall rate of all words that follow, although exactly which words benefit is unpredictable.
3. When looking at all words of a list the filled pauses do *not* consistently affect the overall performance. It is possible that the filled pauses increased the recall rates of words that followed them while at the same time decreasing the recall rates of words that preceded them.
4. The filled pauses have a positive effect that could not be reproduced with the artificial acoustic signal. The effects of the latter seem to be inconsistent leading to the conclusion that the signal is interpreted *differently* by subjects than the filled pause. This difference shows filled pauses to be a *non-obtrusive* element in speech.

The effect of mitigation was tested using adjacency pairs. The only variation in responses was in the usage of filled pauses. The findings can be summarized as:

1. On average, subjects judge responses containing filled pauses to be *friendlier* than those that do not.
2. The mitigation effect in computer generated speech can be considered similar to that of human speech, because in a direct comparison between human and synthetic speech with identical contents the majority of subjects had *no* preference. Additionally the preference between responses with and without a filled pause in human speech mirrored the results of the synthetic speech.
3. Filled pauses can *positively* affect the perceived honesty of a response.

The third part of the experiment asked participants to compare 3 pairs of sentences. Every pair of audio recordings had the same utterance with and without a filled pause. The subjects were asked to state a preference between the two versions in categories of: naturalness, understandability and overall preference. In summary:

1. The overall preference varied between the 3 sentences. The lack of consistency in preference for filled pauses suggests that the preference is dependant on the contents of the utterance.
2. In none of the 3 utterances did subjects significantly judge the filled pause version to be less natural than the unmodified version, suggesting that filled pauses are *not detrimental* to the naturalness of the utterances.
3. Subject preference in the category of understandability is roughly equal to the overall preference for each pair of utterances.

The preference with regards to naturalness as well as the informal testing that preceded the experiment show that the realization of filled pauses through means of the ‘shwa’ phoneme is accurate enough to not produce misinterpretation by unprepared/unprompted listeners.

5.3 Implications

The conclusions that were drawn from the results of the experiment have implications for speech enabled applications. The effects of filled pauses can be used to improve the user experience and the emotional resonance of an agent with the user. Filled pauses don’t only benefit the naturalness of the speech but can also increase the efficiency with which a speech enabled application achieves its goals by manipulating the attention of the listener. In the case of diphone synthesizers the “shwa” phoneme can be used to recreate filled pauses without any additions to the inner workings of the synthesizer. This could extend to other forms of speech synthesis as well. This means that programmers and designers can immediately modify an application to make use of the benefits of filled pauses without updating the speech production. In addition to the findings of this research, the nature of the filled pause accentuates the “written language bias” (O’Connell & Kowal, 2004) in speech synthesis.

In addition to the implications for human computer interaction, the results of the corpus have possible consequences in the field of psycholinguistics. The unexpected link between variation of filled pause duration and the frequency of use was explained using a plausible assertion. However, the underlying cause is uncertain and raises questions about how speakers plan for and realize unlexicalized filled pauses.

5.4 Future Work

This chapter contains suggestions to improve the utilized methods of this research project based on the lessons learned form performing the corpus analysis and the experiment presented in this paper. They are intended to shed light on the parts of the research for which there are still unanswered questions.

The corpus analysis was performed with the specific goal of discovering causal categories and information about the varying pronunciations. The method for collecting the data could be improved upon by doing the following:

- Categorize corpus by parts of speech to discover distribution of filled pauses among adjacent word types.
- Use multiple annotators for measuring filled pause duration to compensate for personal bias.
- Use multiple annotators for categorization.
- Use a corpus with more topical changes, more speakers, all of which native.

The experiment involving the effects of filled pauses was a micro-analysis of an intricate and somewhat unknown process. Improvements on the methods besides a larger number of participants and more limited scope include:

- Use only one filled pause in a free recall exercise, because the effect of a single filled pause lasts for an unknown amount of time.
- Use a different exercise to measure changes in accessibility. The order-of-recall used in the experiment was unable to discover any such change.
- Investigate the effect of filled pauses on other subjective properties besides friendliness and honesty.
- Use a corpus containing more casual and emotional speech.
- Investigate interaction of filled pauses and properties of the speaker (gender, intonation, mood, inherent friendliness) when measuring the mitigation effect.
- Use a large number of varying types of utterances to investigate what sentence types benefit from a filled pause.
- Vary the realization of filled pauses (open/closed, duration etc.) to discover the change in strength of effect.

6 References

Brennan, S. & Williams, M. (1995). The feeling of another's knowing: Prosody and Filled pauses as cues to listeners about the metacognitive states of speakers. *Journal of Memory and Language*, 34, 383-398.

Clark, H. H. & Fox Tree, J. E. (2002). Using uh and um in spontaneous speaking. *Cognition*, 84, 73–111.

O'Connell, D.C. & Kowal, S. (2004). The History of Research on the Filled Pause as Evidence of The Written Language Bias in Linguistics. *Journal of psycholinguistic research*, 33, 459-474

Corley, M., & Hartsuiker, R.J. (2003). Hesitation in speech can... um... help a listener understand. In *Proceedings of the twenty-fifth meeting of the Cognitive Science Society*.

Eakins, B. & Eakins, R. (1978). *Sex Differences in Human Communication*. Boston: Houghton Mifflin

Glanzer, M. & Cunitz, A. R. (1966): Two storage mechanisms in free recall. *Journal of Verbal Learning and verbal Behaviour*, 5, 351–360.

Independent and Dependent Clauses. In *OWL: Handouts: Grammar, Punctuation, and Spelling.* Retrieved on August 26, 2006 from http://owl.english.purdue.edu/handouts/grammar/g_clause.html

Jurafsky, D. & Martin, J.H. (2000) *Speech and Language Processing: An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition*. New Jersey: Prentice Hall

Kowal, S. & Bassett, M.R. & O'Connell, D.C. (1985). The spontaneity of media interviews, *Journal of psycholinguistic research*, 14, 1-18.

Loos, E.E. & Anderson, S. & Dwight H. & Day, Jr. & Jordan P.C. & Wingate J.D. (5-Jan-2004). What is an adjacency pair? In *Glossary of Linguistics Terms.* Retrieved August 26, 2006, from http://www.sil.org/LINGUISTICS/GlossaryOfLinguisticTerms/WhatIsAnAdjacencyPair.htm

Loos, E.E. & Anderson, S. & Dwight H. & Day, Jr. & Jordan P.C. & Wingate J.D. (27-Jan-2004). What is an utterance? In *Glossary of Linguistics Terms.* Retrieved August 26, 2006, from http://www.sil.org/linguistics/glossaryoflinguisticterms/WhatIsAnUtterance.htm

Mehta, G., & Cutler, A. (1988). Detection of target phonemes in spontaneous and read speech. *Language and Speech*, 31, 135-156.

Murdock, B.B., Jr. (1962). The Serial Position Effect of Free Recall. Journal *of Experimental Psychology*, 64, 482-488.

Nass, C. & Lee, K.M. (2001). Does computer-synthesized speech manifest personality? experimental tests of recognition, similarity-attraction, and consistency-attraction. *Journal of Experimental Psychology: Applied,* Vol. 7 No. 3, 171–181.

Rose, R. L. (1998). *The communicative value of filled pauses in spontaneous speech*. Unpublished master's thesis, University of Birmingham, Birmingham, UK.

Swerts, M. A. & Wichmann & Beun, R. (1998). Filled pauses as markers of discourse structure. *Journal of Pragmatics*, 30, 485–496.

Appendix A

Experiment Database Table Structures

Table: subjects
Contains: user personal info
Columns:
name VARCHAR(255) // not used
email VARCHAR(255) //email address of user
age INTEGER(3) //age group of user where 0=-17, 1=18-30, 2=31-50, 3=41-60 and 4=60+
gender INTEGER(1) //gender of user where 0 = male and 1 = female
country VARCHAR(255) //language dominance of user where 0 = native speaker and 1 = 2nd language
experience INTEGER(1) //experience with TTS where 0 = none, 1 = little and 2 = a lot
version INTEGER(1) //version of the experiment performed. Values 1-3. Version corresponds to "Combination" as mentioned in the report.
total_time INTEGER(10) //time take to complete the experiment used to make estimate after pre-tests.

Table: exp1
Contains: values input during exercise 1
Columns:
subj_email VARCHAR(255) //email value to tie to user information
list INTEGER(1) //which word list answer belongs to
answer_num INTEGER(1) //what number this answer is in the sequence
answer VARCHAR(255) //input (word from list)
elapsed_time INTEGER(7) //time it took to recall and write word since previous word

Table: exp2
Contains: data for exercise 2
Columns:
subj_email VARCHAR(255) //email value to tie to user info
fragment INTEGER(1) //what comparison
question INTEGER(1) //comparison (0) or extra question (1)
answer VARCHAR(255) //answer to comparison or question
elapsed_time INTEGER(5) //not used

Table: exp3
Contains: data for exercise 3
Columns:
subj_email VARCHAR(255) //email to tie to user info
sentence INTEGER(1) //what comparison
category INTEGER(1) //category: naturalness(1), understandability(2), preference (3)
choice INTEGER(1) //user input (choice 1-3)

Appendix B

Welkom

Deze pagina is het startpunt voor mijn taalexperiment. Dit is een onderdeel van mijn afstudeeropdracht. Ik zou het heel erg waarderen als u even de tijd zou nemen om hier aan mee te doen. Het experiment duurt in zijn geheel ongeveer een kwartier, afhankelijk van hoeveel tijd u in het herbeluisteren van bepaalde fragmenten wilt stoppen. Het experiment bestaat uit het beluisteren van geluidsfragmenten en het beantwoorden van vragen daarover. Het is van belang voor de resultaten dat het experiment in één keer afgerond wordt. Op deze pagina kunt u meer informatie vinden over wat u nodig hebt om de applicatie uit te voeren en over de veiligheid en privacy m.b.t. uw persoonlijke gegevens. U kunt de applicatie starten door op de onderstaande knop te drukken.

Minimale Configuratie

- Java 1.4 of hoger
- Geluidskaart en boksen of koptelefoon
- Internetverbinding (tijdens het opstarten en beëindigen van het programma is communicatie met een centrale database nodig)

Veiligheid

Zodra u op de knop drukt download de browser een klein XML bestand (met extensie `*.jnlp`) vervolgens aan Java Web Start gegeven wordt om alle nodige data te downloaden en de applicatie te laden. De bestanden die gedownload worden blijven op uw PC staan voor het geval het programma nog een keer uitgevoerd moet worden. Deze kunt u achteraf m.b.v. de “Java Control Panel” verwijderen als u daar bezwaar tegen heeft. Deze is aan te roepen via `jpicpl32.exe` of `javacpl.exe` afhankelijk van de Java versie.

Het programma zelf heeft maar beperkte toegang tot uw systeem en is veilig uit te voeren. Programma’s die via Java Web Start uitgevoerd worden komen in een zogeheten “sandbox” terecht en kunnen geen ongewenste of gevaarlijke acties uitvoeren tenzij u hier expliciet toestemming voor geeft. Dit laatste gebeurd via een dialoogscherm die Java Web Start zelf genereert, hoewel dit in mijn experiment niet nodig is.

Privacy

U wordt aan het begin van het experiment om wat persoonlijke informatie gevraagd. Dit is nodig voor het analyseren van de data. De persoonlijke informatie zal aan geen enkele instantie doorgegeven worden en zal alleen voor mij persoonlijk bekend zijn. Uw emailadres wordt uitsluitend als identificatie gebruikt en ***u zult geen e-mails ontvangen als gevolg van het meedoen aan dit experiment***.

Appendix C

LIST 1	original	filled pause	signal
cabriolet	18	17	18
fiets	20	14	13
racewagen	8	6	6
auto	14	15	13
motor	7	7	2
vliegtuig	10	4	9
truck	10	11	12
bus	5	7	10
rem	12	12	12
stuur	11	9	12
mercedes	14	14	15
snelweg	14	13	11
	n=21	n=21	n=21

LIST 2	original	filled pause	signal
huis	13	13	14
venster	11	13	14
dak	12	8	10
kamer	13	15	14
deur	9	9	12
trap	11	6	9
stoel	9	16	12
kast	3	6	7
stoep	10	11	11
tuin	9	13	11
hek	11	15	11
plafond	18	17	13
	n=21	n=21	n=20

LIST 3	original	filled pause	signal
artikel	19	12	18
schrijver	16	17	15
verhaal	5	3	5
boek	10	9	11
brochure	7	5	5
gedicht	9	9	7
strip	8	4	4
lied	2	1	1
roman	14	14	11
schrift	6	10	10
krant	17	19	17
papier	20	20	17
	n=21	n=21	n=21

LIST 4	original	filled pause	signal
patroon	8	8	3
stal	1	5	1
spel	6	6	4
draad	2	4	4
mok	3	0	4
plaat	3	2	1
raam	8	12	10
plastic	12	8	10
ventiel	14	13	16
broer	5	8	8
poster	13	12	13
lamp	15	14	16
	n=21	n=19	n=21

LIST 5	original	filled pause	signal
kronkel	11	8	10
kooi	6	6	5
tak	7	8	3
berg	11	3	5
potlood	8	7	7
rook	3	4	3
bureau	8	5	7
scherm	4	1	6
handvat	3	5	6
legende	10	10	7
paard	13	20	14
winkel	16	18	17
	n=20	n=21	n=21

LIST 6	original	filled pause	signal
pion	9	9	11
kameel	15	10	15
ruimte	6	8	6
sneeuw	5	8	5
horloge	3	4	3
vuist	5	4	1
slijm	6	2	2
jaar	2	3	5
doorn	1	2	0
afslag	14	16	13
foto	14	14	14
broek	17	17	15
	n=21	n=20	n=20

This is a table with word counts for exercise 1. Words in bold are signaled words. Below each column is the number of times that version of the list was done.

Appendix D

Table of results for exercise 2 of the experiment.

Ratings of the friendliness of the answers in adjacency pairs.

TTS vs TTS+FP	veel minder (in pct)	Minder (in pct)	Gelijk (in pct)	Meer (in pct)	veel meer (in pct)	Num of participants
AdjPair1	0	7	21	63	7	41
AdjPair2	0	9	16	69	4	42
AdjPair3	4	21	48	24	0	41
Human vs Human+FP	veel minder	minder	gelijk	meer	veel meer	Num of participants
AdjPair1	0	9	47	38	4	21
AdjPair2	0	10	30	50	10	20
AdjPair3	4	38	38	19	0	21
Human+FP vs TTS+FP	veel minder	minder	gelijk	meer	veel meer	Num of participants
AdjPair1	4	19	61	9	4	21
AdjPair2	0	42	57	0	0	21
AdjPair3	0	25	35	40	0	20

Ratings of the honesty of the answers.

Honesty	No FP (in pct)	FP (in pct)	Neutral (in pct)	num of ratings
AdjPair1	80	4	14	41
AdjPair2	47	2	50	42
AdjPair3	78	14	7	41
Human Honesty	No FP (in pct)	FP (in pct)	Neutral (in pct)	num of ratings
AdjPair1	80	4	14	21
AdjPair2	40	5	55	20
AdjPair3	76	19	4	21

Appendix E

Results of exercise 3

Sentence 1	No FP	Neutral	FP	Num. of results
Naturalness	27	12	23	62
Understandability	21	17	24	62
Preference	28	12	22	62

Sentence 2	No FP	Neutral	FP	Num. of results
Naturalness	28	4	30	62
Understandability	43	15	4	62
Preference	35	11	13	59

Sentence 3	No FP	Neutral	FP	Num. of results
Naturalness	13	16	33	62
Understandability	16	17	29	62
Preference	13	16	33	62

Subjects chose one of the two audio fragments which they rated higher in the given category or choose to have no preference. One of the audio fragments contains a filled pause and the other does not. The last column represents the total number of entries in that category for that specific sentence.