Universiteit Twente
faculteit der
informatica

Generating natural narrative speech for the Virtual Storyteller

M.Sc. Thesis, March 2004

Human Media Interaction Group
Department of Electrical Engineering, Mathematics and Computer Science
University of Twente, Enschede
The Netherlands

K.J. Meijs

Committee

dr. M. Theune
dr.ir. H.J.A. op den Akker
dr. D.K.J. Heylen
dr. R.J.F. Ordelman

Preface

This thesis is the result of the research I have conducted at the Computer Science faculty of the University of Twente. The last nine months I have worked on a project aimed at the generation of narrative speech. Those aspects of speech that make narrative speech sound as such are among others the prosody of the speech, which was the main area of focus during the project. The project had a lot of variation and covered several different research aspects from speech analysis to experimental evaluation, and last but not least contained some implementation.

I would like to thank several people for their assistance and guidance during the project. First of all my supervisor Mariët Theune for her support and several useful discussions on the subject. My gratitude also goes to the rest of my graduation committee, Dirk Heylen, Rieks op den Akker and Roeland Ordelman for their advice and support. Finally I would like to thank my girlfriend Kim and my parents who have been supportive from the beginning to the end of the project.

Koen Meijs
March 29, 2004

Abstract

If a storyteller recites a story he will try to liven up his speech by using certain rhetorical techniques, such as usage of a specific narrative speaking style and emotion display. Subject of this thesis is the generation of narrative speech. Some important aspects of narrative speech will be analysed and translated to rules which can be used to convert neutral speech into narrative speech. Based on this rules a module is implemented which automatically generates narrative speech based on a certain input text.

The module can be used in the Virtual Storyteller Project, a multi-agent story generation system in which plots are automatically created, converted to natural language and presented by an embodied, virtual storyteller using spoken language with appropriate prosody and gestures. The implemented module is used to pronounce the plots that are created by the virtual storyteller in a narrative way.

Samenvatting

Als een verhalenverteller een verhaal voordraagt, dan zal hij proberen zijn spraak tot leven te brengen door gebruik te maken van bepaalde retorische technieken, zoals gebruik van een specifieke vertellende spreekstijl en weergave van emotie. Onderwerp van dit afstudeerproject is het genereren van verhalende spraak. Enkele belangrijke aspecten van verhalende spraak zullen worden geanalyseerd en vertaald naar regels die gebruikt kunnen worden om neutrale spraak te veranderen in verhalende spraak. Gebaseerd op deze regels is een module geïmplementeerd die op basis van een bepaalde invoertekst automatisch verhalende spraak genereert.

De module kan worden gebruikt in het Virtuele Verhalenverteller Project, een multi-agent verhaalgeneratie systeem. Hierin worden automatisch gecreëerd, vervolgens geconverteerd naar natuurlijk taal en gepresenteerd door een embodied agent, de virtuele verhalenverteller. Deze gebruikt gesproken taal met de juiste prosody en gebaren. De geïmplementeerde module wordt gebruikt om de plots die door de verhalenverteller gegenereerd worden op verhalende wijze uit te spreken.

Table of contents

1 Introduction

1.1 Project description and goals

Several kinds of monologues exist, for example reading the news or reciting poetry. Another type of monologue is recited by a speaker that is telling a story, which is different because he will try to shape the story in such way that it is displayed as expressive as possible. For this he uses common rhetorical techniques such as prosody, facial expressions and gestures. Prosody can have various communicative functions like language act, tension expression, emphasis and emotional expression. These functions manifest themselves in the prosodic features of speech such as melody, rhythm and tempo. In this project we want to find out how a storyteller uses prosody for the purpose of narrative expression.

This graduation project is part of the Virtual Storyteller project, which is a multi-agent story generation system in which plots are automatically created, converted to natural language and presented by an embodied, virtual storyteller using spoken language with appropriate prosody and gestures [8]. The part of the Virtual Storyteller in which our project is involved is the speech output of the virtual storyteller. In order to let the virtual storyteller sound like a real storyteller we need to supply his speech with suitable prosody. In this project we will find out what prosody is appropriate for a storyteller.

Our project consists of two parts. In the first part we will examine which aspects of prosody are of importance in speech in general, besides we will be focus on the specific prosodic functions that are important in narrative speech and what influence they have on the acoustic properties of the speech signal. To perform this investigation, work that has been done before will be examined and narrative speech will be analysed. The goal of this phase is to construct a set of conversion rules that can be applied to neutrally spoken speech and convert the input speech to narrative speech. This conversion is based on prosodic functions that are important in storytelling, such as tension expression and emphasis.

The second part of the project follows a more practical approach and comprises the implementation of a system that uses the rules that were formulated in the previous phase to automatically convert textual input to narrative speech. For this purpose we will make use of a currently existing Dutch text-to-speech engine. During the entire project the Dutch language is used as base language for both the examination and synthesis of narrative speech.

The steps that are involved in the automatic generation of narrative speech by our system are as follows (see figure 1.1, grey boxes are the processes that are performed by the system, the white boxes represent data).

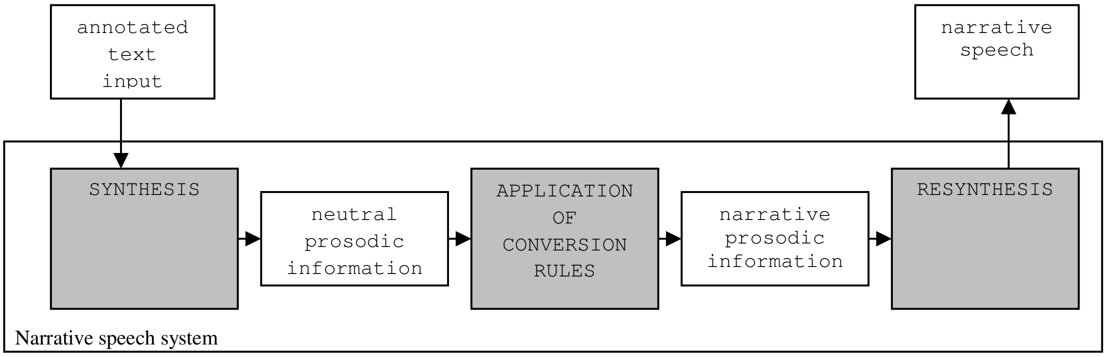

annotated text input
narrative speech
SYNTHESIS
neutral prosodic information
APPLICATION OF CONVERSION RULES
narrative prosodic information
RESYNTHESIS
Narrative speech system

Figure 1.1. Steps taken in automatic generation of narrative speech

First the input text that has to be spoken is written down in a mark-up language that allows the annotation of important prosodic functions of storytelling. Then the text is synthesised by the text-to-speech module without applying any modifications to it, so the result is a neutrally spoken version of the text. This output of the synthesis contains information about prosodic features that are of importance, our conversion rules will be applied to this prosodic information resulting in narrative prosodic information. Next step is to resynthesize the text based on the narrative prosodic information, resulting in narrative speech.

1.2 Report outline

In this report all phases of the project are described chronologically. First underlying and related theory about emotional speech will be explained in the theory section (chapter 2). In this section we determine which prosodic functions are important in storytelling and which acoustic features of speech can realise these functions.

Next step is to perform an analysis of narrative speech, consisting of analysis of narrative speaking style (chapter 3) and analysis of tension course (chapter 4). The goal of this phase was to determine which acoustic properties of speech are responsible for the presence of narrative style in speech and in which degree they are influenced by narrative style. Based on the results of the analysis a conversion rule model is set up which can be used to transform a neutral speech fragment into a narrative speech fragment (chapter 5).

Before we can use these rules in an implementation we have to verify their correctness. This is done in an evaluation in which a listener experiment is conducted. The general experimental setup of the successive evaluations is first provided (chapter 7). The two following chapters describe two successive evaluations, the constant evaluation (chapter 8) and the conversion rule evaluation (chapter 9).

After evaluation of the conversion rules they are used in the implementation, in which a software module is built that implements the conversion rules. This module can create narrative speech using an annotated text as input (chapter 10). The last step in the project is to evaluate the quality of storytelling of the implemented module (chapter 11). This report ends with a conclusion and recommendations (chapter 12 and 12).

2 Theory

2.1 Introduction

Communication involves not only spoken words, but includes linguistic elements, paralinguistic elements and non-verbal elements like facial expression, co-speech gestures and non-speech sounds, which all contribute some meaning to the speech signal. Although those non-verbal elements are of great importance in storytelling as well, for the purpose of generating narrative speech we first need to focus on the linguistic and paralinguistic elements that are involved.

In this section we look at comparable work that has been done already to get an idea of the current state of affairs of emotional speech synthesis. Several aspects of emotional speech will be described here from the perspective of storytelling.

The following subjects will be treated in this section:

- Narrative speech
- Prosody
- Paralinguistics
- Generation of narrative synthetic speech
- Text-to-speech

This section provides a wide view on the subject area of the project. Because of its magnitude not all aspects of emotional speech that are discussed in this section will be used in the project. Based on the theory that is discussed in this section, a selection of prosodic and paralinguistic functions is made which will be the aspects of emotional speech that will be included in this project.

2.2 Narrative speech

Narrative speech is a form of speech that in great extent depends on the rhetoric of the speaker. Rhetoric is the effective use of language aimed at *convincing* the listener [2]. In the case of a storyteller, convincing must be seen as the process of creating empathy and sympathy from the listeners for the characters and event in a story.

A narrator uses several techniques to realise this:

- *Specific narrative style*[1]
  A narrator who tells a story uses a completely different speaking style than a newsreader. A storyteller will use certain tempo and pitch variations to emphasise and clarify certain elements of the sentence that are important. This technique is especially observable in the case of child stories.
- *Presence of emotion and attitude*

[1] In this document we will use narrative style to denote the specific storytelling speaking style a speaker is using. Narrative speech is defined as speech that includes that style together with emotion, tension etc.

The presence of emotional or attitudinal expression in the speech based on the plot increases the engagement of the listener of the story.

- *Tension course*
  Every story contains a certain dramatic tension[2] based on the events that take place in the story. Depending on this tension course the speaker creates silences or evolves towards a climax. When for example a sudden event takes place, the narrator communicates the tension change that is involved in this event in his speech to engage the listeners.
- *Use of voices for characters*
  A narrator can use various voices in order to realise a distinction among characters in the story (for example a witch with a high grating voice).

If we want to create realistically sounding narrative speech, we should include all techniques that a real life storyteller uses in the creation process. Most important of those techniques is the narrative speaking style, because this technique has a continuously present effect on the speech. The other techniques (emotion and attitude, tension course and voice use) have an incidental effect, meaning they are only included when needed in the story. Therefore they are not essential but do contribute a lot to the naturalness of the speech.

All of the above techniques fall under prosody or paralinguistics, therefore in the next paragraphs these two subjects will be explained in more detail.

2.3 Prosody

In [1] prosody is defined as the whole of properties of a speech utterance that cannot be reduced to the succession of phonemes (vowels and consonants). A phoneme is the smallest speech sound unit that can be marked off by time. A phoneme encloses all possible pronunciation variations of a certain speech sound.

In general prosody is used to refer to variations in pitch, intensity, tempo and rhythm within an utterance, which are determined by prosodic functions accentuation and phrasing. Accentuation is the emphasising of certain words in an utterance by changing the pitch while pronouncing them. Phrasing is the division of utterances into intonational phrases, often separated by a pause, rise in pitch and lengthening of pre-boundary speech sounds [30].

In [1] the following communicative functions of prosody are enumerated:

- *lexical function*
  The lexical function realises the distinguishing of word forms by means of differences in melody and/or accent. Although especially vowels and consonant are responsible for this prosody also plays a role in this. One example of lexical function of the prosodic accent is the appearance of the accent in the Dutch word ‘kanon’, which has two different meanings depending on the accent’s position. The lexical function realized by melody is observable the Chinese language, in which identical sequences of phonemes can have different meanings depending the melody that is used to pronounce them.
- *phrasing function*

[2] Tension is defined as thrill or excitement that is present in the story.

The phrasing function realises the dividing the speech stream into information units by means of variation in tempo and use of pauses.

- *information structuring*
  Inside marked off units prosody functions in two ways:
  - attentional: Denote with the help of accenting which unit is communicatively more or less important.
  - intentional: Speaker adds nuancing to the information structure. Every language has a limited set of rising and falling pitch movements. By introducing an alternative pitch movement a speaker can realise a nuancing of meaning, for example in a facetious remark.
- *attitude signalling*
  What is the opinion of the speaker with regard to the verbal content of the sentence
- *emotion signalling*
  In which state of mind is the speaker

These last two functions have a lot of overlap and in literature they have been interpreted in different ways. In [3] a plausible distinction is made; attitude is interpreted as the categorisation of a stimulus object based on an evaluation dimension. So an attitude is the psychological tendency to evaluate an entity by attributing a certain amount of approval or disapproval to it.

Emotions are conceived as discrete states of mind like for example ‘angry’ or ‘sad’. A common approach for the description of emotions is the division of these in primary and secondary emotions. Just like primary and secondary colours, primary emotions form the base emotions and secondary emotions are formed by mixing those. Usually as primary emotions the ‘big six’ which originate from the Darwinian perspective [4] are distinguished: fear, anger, happiness, sadness, surprise and disgust. According to Darwin these emotions represent survival-related patterns of responses to events in the world that have been selected for over the course of our evolutionary history. Therefore these emotions are considered fundamental or primary and all other emotions are thought to be somehow derived from them.

In prosody we distinguish on one hand above mentioned information entities or functions that are fulfilled by prosody, on the other hand we have the prosodic features, which are the specific form phenomena through which the prosodic functions are communicated. The most important from phenomena are the base frequency, intensity and temporal structure of the signal. The base frequency is the repeat frequency of base period of the sound signal, corresponding with the observed pitch. The intensity or loudness is the volume of the sound. The temporal structure of the sound can be split up in three acoustic features: overall tempo, pausing and the duration of local units (for example vowels).

2.4 Paralinguistics

Besides prosodic features a speech signal also contains so called paralinguistic features. Paralinguistic features are vocal effects that are primarily the result of physical mechanisms other than the vocal cords, such as the direct result of the workings of the pharyngeal, oral or nasal cavities. These features are determined by for example the age or gender of the speaker.

In [5] there is distinction between *voice qualities* and *voice qualifications*. Voice qualities include features of more or less continuous nature and there exist shortly appearing features. The continuous features are pronunciation styles like whispering, creaky or breathy speech. The intermittently appearing features are other vocal effects that are caused by fluid control and respiratory reflexes like clearing of throat, sniff, gulp and yawn. Under voice qualifications terms like laugh, cry and tremulous voice are interpreted.

In the case of storytelling paralinguistic features are used in the storyteller's speech to realise certain characters that occur in the story, for example a witch with a high creaky voice.

2.5 Generation of narrative synthetic speech

The preceding two paragraphs describe prosodic and paralinguistic functions that are used in speech in general. These functions are present in both neutral[3] and narrative speech, but the acoustic realization of them can differ among the two speaking styles. The most differing functions are most interesting to examine, because they realize what narrative speech distinguishes from normal speech. So if we want to produce narrative synthetic speech we should study those prosodic and paralinguistic functions that make the difference and realise these functions in the speech synthesis.

In the preceding paragraph about prosody (§2.3) five functions of prosody have been enumerated. One of these functions that has a distinguishable effect in narrative speaking style is the attentional information structuring function. This function is applied if a storyteller uses accenting to increase the relevance of certain sentence parts with respect to other parts. Another prosodic function that is of significance in storytelling is the addition of expressiveness. If affect or emotion is conveyed in speech, this will make a narration more natural and bring it to life.

As in neutral speech we assume that the speaker constantly uses a neutral voice, it is clear that by varying paralinguistic features voice quality and voice qualification during storytelling, an effect is realized that contributes to the quality of narrative speech.

Summarizing, for the realization of narrative speech the most distinguishing prosodic functions are the attentional information structuring and the attitude and emotion signalling, while all paralinguistic features are of distinguishing contribution to narrative speech. So in the generation of narrative synthetic speech those functions are of importance.

2.6 Text-to-speech

2.6.1 Text-to-speech techniques

In [6] three shortcomings of synthesised speech with respect to human speech are given: insufficient intelligibility, inappropriate prosody and inadequate expressiveness[4]. Intelligible

[3] Neutral speech is defined as speech that contains no expressiveness or specific speaking style, but is spoken as neutral as possible.

[4] In this article prosody is considered not to include emotional and attitudinal functions

speech is essential because otherwise words cannot be recognised correctly. Without appropriate prosody the function of prosody clarifying syntax and semantics and aiding in discourse flow control is not fulfilled. Inadequate expressiveness results in lack of providing any information about the mental state and intent of the speaker. For the purpose of creating narrative synthetic speech, those three aspects are of equal importance. The intelligibility of narrative speech is an aspect that depends on the text-to-speech technique that is used, which we will discuss here.

There exist several different techniques that can be used for the generation of synthetic speech. When we want to create narrative synthetic speech several parameters like pitch and voice quality must be adaptable. The existing techniques provide control over these parameters to very different degrees [7], which means not all techniques are in equal degree suitable to be used for creation of emotional speech. Here we will discuss three well known techniques and look at the advantages and disadvantages of each.

The first technique is *formant synthesis*, also known as *rule-based synthesis*. This technique creates acoustic speech by simulating the human speech production mechanism using digital oscillators, noise sources, and filters; no human speech recordings are used. This approach offers a high degree of flexibility and control over the acoustic parameters related to voice source and vocal tract, which is interesting for modelling emotional speech. A disadvantage of this technique is that is sounds ‘robot-like’ or unnatural compared to other techniques.

Another approach is the concatenative synthesis, which includes *diphone concatenation*. A diphone is a speech unit that starts at the middle of one phone and ends at the middle of the next. In diphone concatenation recordings of human spoken diphones are concatenated in order to create the speech signal. During the syntheses signal processing techniques are use to generate the desired base pitch of the speech. This processing introduces some distortion in the speech signal, but in general the naturalness is better than that of speech created by formant synthesis. Most diphone synthesis systems offer the possibility to adapt pitch, duration and intensity of the signal, but there is no control over voice quality of the speech. Because human recordings of a certain voice are used, only one voice with specific voice quality features can be used in the synthesised speech.

The synthesis technique that is considered most natural is *unit selection*. For this technique a large database of human speech recordings is used (contrary to diphone concatenation in which a relatively small database is sufficient). Out of this database speech units of variable size are selected based on a certain selection process. The selection process may be based on certain parameters like for example pitch and duration. Because of its large database size this technique results in the most natural speech of all techniques, but results can be very bad when no appropriate units are found. A disadvantage of this technique is that for the creation of (emotional) speech a large database of speech is required (speech data for each emotion).

Comparing the three techniques, it turns out that the degree of naturalness of each technique depends on the method of acoustic modelling. Increasing naturalness of the generated speech consequently gives rise to a lower degree of flexibility.

2.6.2 Synthesising algorithms

Two well-known synthesising algorithms that are used for speech synthesis and manipulation are PSOLA and MBROLA [29]. PSOLA performs a pitch synchronous analysis and synthesis of speech, applying pitch and duration manipulations by using a window based on the fundamental frequency for each pulse in the speech signal and adapting the size of the window depending on the desired manipulation. PSOLA delivers high quality speech if the manipulations are kept small and the spectral discontinuities at selected unit boundaries are kept small.

MBROLA avoids concatenation problems by re-synthesising voiced parts of diphones with constant phase at constant pitch, in this way smoothing the unit boundaries if the segments to be concatenated are voiced and stationary at their boundaries. The problem with large manipulations is still present however.

2.6.3 Dutch Text-To-Speech engines

Because in this project the language of the texts that need to be spoken is Dutch we will take a look at Dutch text-to-speech engines that are available. From the perspective of narrative speech it is important that the text-to-speech engine we will use offers sufficient flexibility, so there should be the possibility to control certain acoustic parameters that are essential in narrative speech.

The first text-to-speech engine we have considered is Nextens [31] (‘Nederlandse extensie voor tekst-naar-spraak’). Nextens is an open-source TTS system for Dutch developed by the Department of Language and Speech of the University of Nijmegen and the Induction of Linguistic Knowledge group of the University of Tilburg. For general architecture and basic facilities Nextens relies on the Festival system. It uses the MBROLA diphone synthesiser for waveform synthesis and natural language processing tools for Dutch grapheme-to-phoneme conversion, POS tagging, accent placement and prosodic phrasing. Nextens offers the possibility to process annotated text input, using the SABLE XML standard. Acoustic parameters pitch, intensity and speech rate can be adjusted, both in absolute (by specifying the adjustment by an absolute value in the property’s quantity) as in relative way (by specifying the adjustment by a percentage or a descriptive term like ‘small’ or ‘large’).

Another Dutch text-to-speech engine that exists is Fluency TTS [32], developed by Fluency, a division of Van Dale Lexicografie. Fluency is a commercially available TTS system for Dutch. It uses the MBROLA diphone speech synthesiser for waveform synthesis. Fluency includes the possibility to store prosodic phoneme information in a specific format that allows the notation of absolute pitch and duration values of the phonemes.

After the analysis of narrative speech has been conducted and we know which acoustic features are of importance in the generation of narrative speech, we will make a decision which of the text-to-speech engines to use in our implementation.

2.7 Function selection

The amount of prosodic and paralinguistic functions that can be utilised in storytelling is extensive, so if we want to analyse narrative speech it’s not possible to investigate all functions elaborately. For this reason we choose a subset of functions that we want to examine in our analysis.

In paragraph 2.5 we have already selected those prosodic and paralinguistic functions that are specifically important for narrative speech. Those are the functions that realize the distinction between neutral and narrative speech; prosodic functions like lexical and phrasing function (§2.3) are not included in our study because they are not specific for narrative speech. Another reason why we leave these functions aside is that they have been studied extensively and are implemented in any text-to-speech engine already. So there's no need to examine these functions here.

In this project we will focus primarily on the most important function that is used in narrative speech, the specific narrative style. We will also include the tension course function in our study, but we will not include the addition of emotion/attitude and paralinguistic (voice quality) functions. The reason for leaving out the emotion and attitude function is that in the past there has been a lot of research in the area of affective speech already, resulting in a great amount of guidelines that can be used to add emotion to speech. It is not useful and desirable to reinvestigate this. The voice quality of speech is a separate area of study, which is not included because we want to concentrate on the prosodic functions of narrative speech.

3 Analysis of narrative speaking style

3.1 Introduction

In the theory section (chapter 2) it has been explained which functions of prosody exist and by which acoustic properties of a speech signal they can be realised. The next step is to take a look at the prosodic differences when we compare a neutrally spoken text to a text that is spoken by a storyteller. In the second case in general the narrator will have the tendency to use certain rhetorical techniques to liven up the story that he is telling. Most important of those techniques are the distinct narrative style and the addition of tension course depending on the events that take place in the story. The use of these techniques results in adaptation of the acoustic features of the speech, which are properties of speech that can be measured.

In this analysis we determine which acoustic features are influenced by the use of storytelling techniques by comparing them to the acoustic features of neutral speech. After we have derived how the rhetorical techniques are realised in the acoustic features we will formulate rules which can be used to transform a neutral speech fragment into narrative speech. These rules can later on be used in a text-to-speech module so we can automatically generate narrative speech.

This chapter describes the analysis of the first rhetoric function that is examined, the narrative speaking style. The following chapter (chapter 4) describes the analysis of the other function, the tension course. Each of these chapters consists of a description of the setup of the analysis (§3.2-§3.4 and §4.2), including a description of the prosodic features that will be examined, the analysis material and the analysis tool. The chapters are concluded with a description of the results of the analysis (§3.5 and §4.3) and a summary (§3.6 and §4.4).

Based the outcomes of the analysis of the two functions, conversion rules are formulated which can be used to transform neutral speech to narrative speech. These rules are described in chapter 5.

3.2 Prosodic features

Before we start the analysis we should determine which acoustic features of a speech signal are possibly influenced by the rhetoric function narrative speaking style. Many acoustic features can be distinguished which are possibly influenced by the rhetoric function, but not all of these features are of equal importance. The most important features of the speech signal are the fundamental pitch value and range, intensity, tempo and pausing of the speech signal. All these features will be included in the analysis that is performed. Special attention will be paid to one specific aspect of tempo that seems important in realisation of narrative style, namely the variance in the duration of vowels in words that carry sentence accent. This aspect will be examined as well.

After all features have been analysed we have obtained global statistical information about the acoustic features. Then we have obtained a global image of the difference between the speakers, but it is possible that local differences are observable as well. Because of this we will take a look at the pitch and intensity contour course during the fragments, to see whether certain returning pitch or intensity contour movements can be observed.

3.3 Analysis material

In order to conduct this analysis three kinds of datasets have been used. All speech fragments in the three datasets are in the Dutch language. The first dataset contains samples that are taken from Dutch news broadcastings and will be used as a reference baseline for neutral speech. The second consists of storytelling samples but in this case the stories are adult fairy tales. The third dataset consists of samples taken from children's stories read by a storyteller. For the purpose of the analysis of narrative style we will take a selection of fragments from each of the datasets and statistically analyse their pitch, intensity, tempo, pause and duration properties. A comparison will be made so we can see which features differ among different speakers.

Before this quantitative analysis was carried out, a general qualitative analysis is conducted in which a larger amount of samples of each dataset is analysed globally. This analysis is performed to get a general idea of the behaviour of the acoustic features (§3.5.1).

During the analysis it is important to keep in mind that the fragments are spoken by three different speakers with consequently differences in general speech properties. Every person has a unique way of speaking, for example a person can speak in a very monotonic way (low pitch range) or have a relatively high base speech tempo or pitch. A concrete consequence of this aspect is that we can't compare absolute feature values of the three speakers (which would be the case if the three datasets were spoken by the same person). So if we want to detect differences in pitch or intensity statistics among the speakers we always have to use the average statistics of the speaker in question as a reference for comparison. Statistical measures like the standard deviation and spreading are good examples of trustworthy values for comparison. Another reason why relative statistics are desirable is the effect of local minima and maxima. For example, if we take a look at the *pitch range* it is clear that one extremely high value of pitch results in higher pitch range although this raise is only caused by an incidental value, so it can be dangerous to use the range in our comparison.

In the selection process of the fairy tale fragments one important requirement is that all fairy tale fragments that are selected may not contain any evident emotions or voice transformations (which are frequently used to express a story-character). The fragment must only contain parts of the text in which the narrator speaks in indirect speech (only fragment Adult_2 contains an utterance in direct speech, but it is spoken in the narrator's neutral voice). So we have selected the most neutral speech of a fairy tale narrator. In this way we avoid that other rhetoric functions influence our analysis results in a negative way. The fragments that are used in analysis are given in appendix A.1. We will evaluate a total of eight fragments, three newsreader fragments of respectively 20,0, 3,1 and 6,5 seconds, three child storyteller fragments of respectively 6,3, 7,0 and 4,1 seconds and two adult storyteller fragments of respectively 5,1 and 9,6 seconds.

3.4 Analysis tool

In order to analyse speech signals we will use a tool called *Praat* [33]. This tool is able to perform temporal, spectro, formant, intensity and pitch analysis. For our purpose there exist more suitable analysis tools with usually similar possibilities, but we chose for *Praat* because this tool offers very elaborate process and analysis possibilities, showing results in well-organised way. Besides *Praat* also offers sound manipulation facilities, which are very useful in the subsequent evaluation phase of the project, in which manipulation speech fragments will be created.
Figure 3.1 shows a fragment of utterance *News_2* as it looks in *Praat*, figure 3.2 shows fragment *Child_1*, which is pronounced by a storyteller.

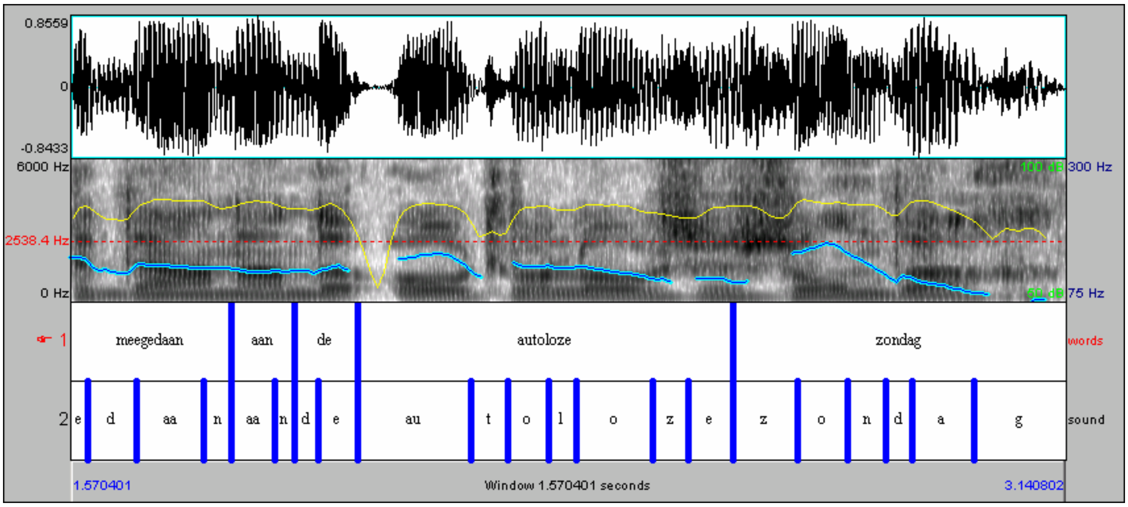


Figure 3.1. Newsreader speech sample in *Praat*

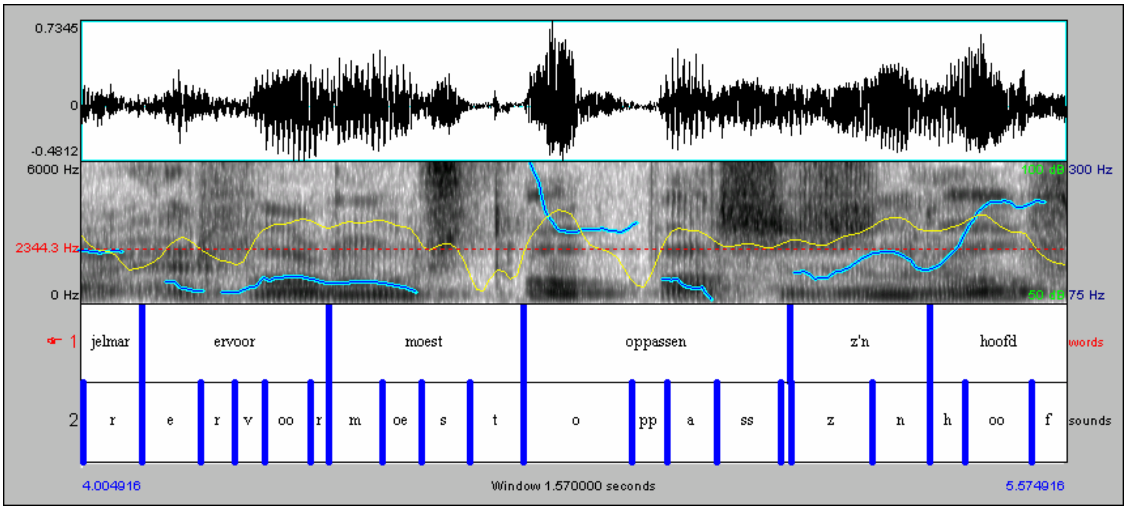


Figure 3.2. Storyteller speech sample in *Praat*

In both figures the upper signal is the waveform speech signal itself. The bar underneath contains the spectrogram, the pitch contour (dark line) and the intensity contour (white line). The lowest two bars show the utterance segmented in words and phones.

3.5 Analysis results

3.5.1 Global analysis

When making a global comparison of the two fragments, at first sight it is visible that the newsreader has a more constant level of intensity and pitch than the child storyteller, who has much more variation in his pitch and intensity levels. This observation indicates that the nature of the text determines a certain narrative style resulting in differences in the acoustical properties of the signal.

When comparing the pitch and intensity contours of the newsreader and the adult storyteller, we see that the adult storyteller has more varying contours with respect to the rather constant contours of the newsreader, but the difference between those two speakers is much smaller than that between the newsreader and child storyteller. Because we want to find out which features of the neutral speech and the narrative speech are most differing (§2.5), in the analysis we will primarily focus on the newsreader and child storyteller, although the adult storyteller will be analysed as well to show this observation by quantitative analysis (for pitch, intensity and tempo).

3.5.2 Pitch analysis

We have used the following statistical properties in the pitch analysis: frequency minimum, maximum, average, range and standard deviation. We also included the 90%-quantile, the frequency value below which 90% of the values is expected to lie. This value gives a more normalised estimate of the maximum values of the signal (no influence of local maxima).

We measure the spreading by subtracting the 10%-quantile from the 90%-quantile. In this way we know the bandwidth in which 80% of the values are lying. So any values in the first 10% and the last 10% of the bandwidth are thrown away, in this way we exclude any extreme values.

The last property we examine is the mean absolute slope of the pitch contour, which indicates the average speed of rise and fall of the pitch.

The results of the analysis are shown in table 3.1.

fragment name	**min (Hz)**	**max (Hz)**	**range (Hz)**	**average (Hz)**	**standard deviation (Hz)**	**90%q (Hz)**	**90%-10% (Hz)**	**mean abs. slope (Hz/s)**
News_1	76,1	296,5	220	119,6	21,1	149,1	50,9	238,3
News_2	86,9	181,4	94,5	130,4	21,0	163,2	55,3	255,6
News_3	83,9	188,5	104,6	121,6	20,0	148,2	48,9	234,9
average				123,9	20,7	153,5	51,7	242,9
Child_1	76,4	367	291,6	134,0	46,4	195,1	105,0	366,6
Child_2	75,0	395,8	320,9	123,8	46,3	177,6	90,3	311,7
Child_3	74,9	228,2	153,3	106,8	30,9	141,8	64,3	296,2
average				121,5	41,2	171,5	86,5	324,8

Adult_1	75,0	186,5	111,5	123,2	23,3	156,3	57,2	192,3
Adult_2	93,8	252,7	158,8	124,5	25,4	152,4	52,0	170,4
average				123,9	24,4	154,4	54,6	181,4

Table 3.1. Analysis results for narrative style

As expected, the absolute values like min, max and range are quite divergent for different fragments of the same speaker so they don't give us a solid base to build on. On the other hand, the spreading and slope values do give us indications of how the speaking style changes depending on the text type.

The biggest difference is observable between the values of the child storyteller and the newsreader. When comparing these, in the storyteller's case there is an average increase of 50% of the standard deviation in relation to the newsreader, and the 90%-10% range is increased with 67%. The mean absolute slope of the pitch contour is increased by 34%. This means that the pitch values of the storyteller are more spread and that the slope of the pitch contour is steeper.

When we compare the adult storyteller to the newsreader, there is not much difference perceptible. There is a slight increase in standard deviation and 90%-10% range, but this difference is so small it cannot be attributed to specific narrative style, because this could very well be the personal speaking style of the speaker. Besides, the adult storyteller has a lower mean absolute slope than the newsreader, so the speaker's pitch contour has a more horizontal form than the others.

Taking a look at the pitch contours of the three speakers (dark line in fig. 3.3, fig. 3.4 and fig. 3.5), these observations can be read from the diagram as well. All fragments have approximately a length of 3 seconds. It is clearly visible that the pitch contours of the newsreader and adult storyteller resemble, but that the child storyteller's contour has much more peaks and looks freakishly.

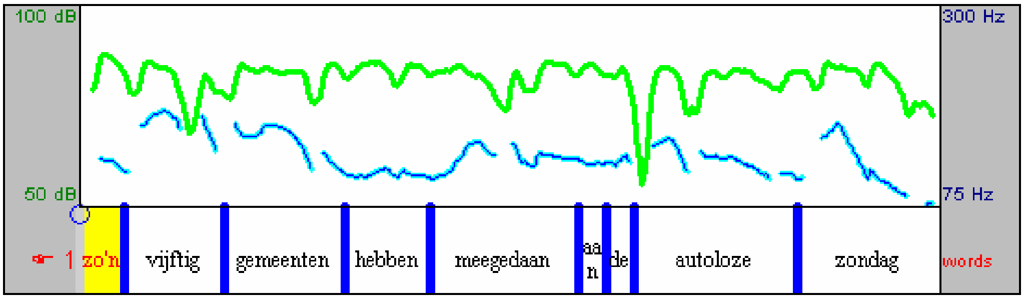


Figure 3.3. Newsreader pitch and intensity contour

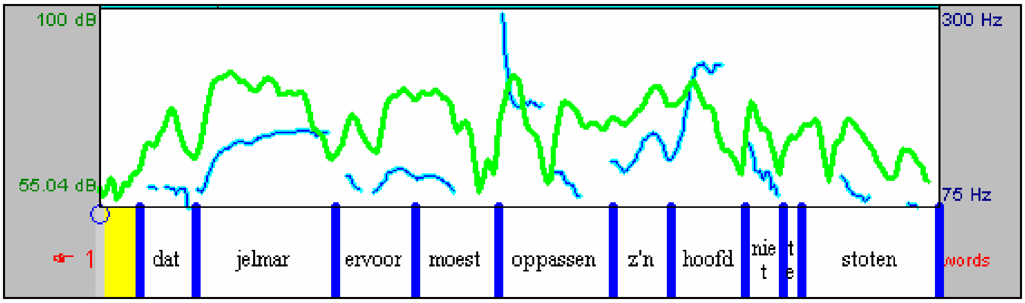


Figure 3.4. Child storyteller pitch and intensity contour

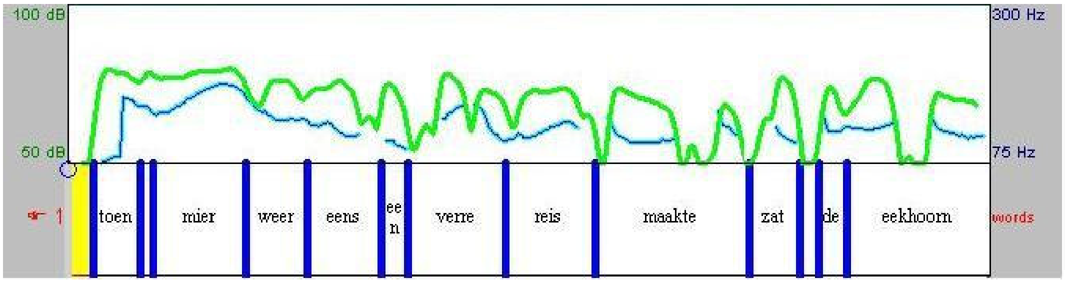


Figure 3.5. Adult storyteller pitch and intensity contour of utterance
'Toen de mier weer eens een verre reis maakte zat de eekhoorn ..'

3.5.3 Intensity analysis

Before we can perform the analysis of the intensity the fragments have to be pre-processed. The reason for this is that *Praat* includes silence (pauses) in the calculation of the intensity statistics, resulting in distortion of statistical information. So we remove silences that occur at the end of utterances from our signals, because we only want to use statistics of the speakers while speaking, not while pausing (pausing is presumed to be influenced by narrative style, so including pauses may distort the results).

The following table (table 3.2) shows the average and the standard deviation of the intensity of the fragments.

name	**average (dB)**	**SD (dB)**
News_1	79,9	5,8
News_2	82,1	3,1
News_3	80,7	4,8
average	80,9	4,6
Child_1	72,5	6,5
Child_2	68,6	9,0
Child_3	68,1	7,2
average	69,7	7,6
Adult_1	66,0	9,8
Adult_2	61,8	11,5
average	63,9	10,7

Table 3.2. Intensity values

The average intensity of the newsreader turns out to be higher than the averages of the storytellers. Speech intensity can be very divergent among different speakers depending on the person itself. For this reason we will not draw any conclusions from the average intensity results. Moreover, another possible cause for the varying results may be that the sound signals aren't recorded at the same intensity.

Still, we can take a look at the standard deviation, which provides us a reliable measure for the spreading. The standard deviation of the child storyteller is 65% higher than the newsreader; the standard deviation of the adult storyteller is 132% higher than the newsreader. This means that the two storytellers have more varying intensity values than the newsreader.
This is also visible in the intensity contours of the three speakers (light grey line in figure 3.3, 3.4, and 3.5). It is remarkable that the newsreader keeps a very constant level of intensity during his utterances, while both storytellers have much more variation in their intensity levels.

3.5.4 Tempo

A speaker has the natural tendency to make all syllables have equal duration [1]. Therefore the number of syllables per second can be used as a measure to compare speech tempo. The following table (table 3.3) lists the average speaking tempo in syllables per second for the three speakers. We have analysed about five sentences in a row for each speaker (see appendix A.2 for the analysed fragments and the exact length of pauses). In the results we distinguish between the tempo with and without pauses that occur between two sentences. The latter is a measure that can be used later on in the interpretation of vowel duration results.

Speaker	tempo in sps (no pauses)	tempo in sps (including pauses)
Newsreader	6,46	5,77
Child storyteller	3,50	3,04
Adult storyteller	4,77	3,63

Table 3.3. Average speaking tempo

Looking at the general tempo (including pauses), it is clear that the newsreader has a faster speech rate than the storytellers. The newsreader speaks 89% faster than the child storyteller; the newsreader speaks 59% faster than the adult storyteller.

3.5.5 Pausing

We distinguish two kinds of pauses. The first is the pause that occurs between two sentences (outside pause), the second the pause that occurs inside the sentence itself (inside pause), usually in the case of a subordinate clause or a sentence in indirect speech.
We have measured the duration of four fragments of inside pauses and four fragments of outside pauses of the newsreader, the same goes for the child storyteller. Table 3.4 lists the average duration of the fragment combinations of pauses and speakers.

pause	duration newsreader (sec)	duration child storyteller (sec)
inside	0,324	0,451
outside	0,527	1,303

Table 3.4. Average pause duration

It is evident that the storyteller takes longer pauses than the newsreader both inside and outside the sentence. The inside pause of the child storyteller is 39% longer than the pause of the newsreader. This percentage is 147% for the outside pauses. These results correspond with the pause duration values determined in [9] for child stories and news readings.

3.5.6 Duration of vowels in stressed words

In every sentence one or more words carry a sentence accent. This means they play a key role in the meaning of the sentence and as a consequence earn extra focus. In order to find out whether the duration of stressed vowels in words with sentence accent is influenced by the narrative style, we have compared the duration of several vowels spoken by the newsreader and the child storyteller. The reason we have excluded the adult storyteller is that if there would exist any difference in vowel duration, this is most evidently visible in the comparison of the newsreader and child storyteller, because they have the most distinct narrative style compared to each other. The following tables list the average duration of regularly occurring accented vowels of both speakers (the speech fragments where these vowels come from can be found in appendix A.3, including the vowel duration measurements) and the average taken over all analysed accented vowels. Table 3.5 contains values for the long vowels (e:, o:, a:)[5], table 3.6 for the short vowels (E, O, A, i.):

	newsreader		**child storyteller**	
vowel (SAMPA)	**average duration (sec)**	**standard deviation**	**average duration (sec)**	**standard deviation**
e:	0,136	0,018	0,145	0,039
o:	0,138	0,021	0,182	0,007
a:	0,119	0,037	0,151	0,034
average	**0,131**		**0,159**	

Table 3.5. Average duration of long accented vowels

	newsreader		**child storyteller**	
vowel (SAMPA)	**average duration (sec)**	**standard deviation**	**average duration (sec)**	**standard deviation**
E	0,078	0,015	0,086	0,020
O	0,066	0,013	0,112	0,051
A	0,087	0,018	0,103	0,047
i.	0,087	0,006	0,098	0,013
average	**0,080**		**0,100**	

Table 3.6. Average duration of short accented vowels

[5] Here the SAMPA [1] notation for Dutch is used

In [1] it is stated that the duration of long vowels is approximately two times as long as the duration of short vowels. The speakers don't completely apply to this assertion with average values for the newsreader and child storyteller of 0,131 and 0,159 seconds for long vowels against 0,080 and 0,100 for short vowels. The duration differences between long and short vowels in our measures are smaller than those in literature, but still long vowels are considerably longer than short ones.

What can be concluded from these results from the perspective of narrative style is that looking at the average values it is not possible to observe any direct correlation between vowel length and narrative style. The average duration of all vowels spoken by the storyteller is longer, but we have to take into account that this may also be caused by the fact that the overall speech tempo of the storyteller is slower than the newsreader's. In the *tempo* section it turns out that the newsreader speaks 85% faster than the child storyteller if we remove outside pauses. This means that the average duration of an utterance of the child storyteller is 85% longer than that of the newsreader. The difference in average vowel duration between the two speakers is smaller (20%-25%) than the difference in tempo, so the vowel duration difference must be attributed to the general tempo.

Although we can't say anything based on the average duration values, when we look at the standard deviation of the duration values, it turns out that the duration values of the storyteller have about equal or higher standard deviation. So there turns out to be quite some variance among the occurrences of the same vowel.

This takes us to another approach concerning duration of vowels; it is possible that the same vowel takes on different durations depending on its part of speech. We will now perform a small analysis based on this assumption (to show the duration differences among different part of speech categories we will not only list accented vowels but all vowels in the fragment).

We will examine this by calculating the relative duration of vowels of a sentence fragment spoken by a newsreader and a sentence fragment spoken by a storyteller. The newsreader fragment was not selected for a special reason; the storyteller fragment contains some words that are suspected to be increased in duration because of their part of speech nature (the adjective part "zo laag"). The sentence fragments are the following, the absolute vowel durations can be found in appendix A.4:

Newsreader: *"Zo'n vijftig gemeenten hebben meegedaan .."*[6]
Storyteller: *".. liepen door een lange gang die zo laag was dat Jelmar .. "*[7]

In order to calculate the relative duration, we first calculate the tempo of both speakers in syllables per second (4,3 s.p.s. for the storyteller, 6,0 s.p.s. for the newsreader). Then we measure the absolute length of the vowels. To be able to compare them we should multiply the fastest reader's duration values by the quotient of the newsreader's tempo and the storyteller's tempo (in fact we now slow down the tempo of the newsreader's speech, making it comparable to the storyteller's speech) resulting in normalised vowel durations.

[6] Fragment from Dutch Radio 1 News, September 21, 2003, 18.00u, spoken by Onno Duyvené de Wit, male newsreader

[7] Fragment from "Klaas Vaak", male storyteller, , "Luister sprookjes en vertellingen", Lekturama.

To calculate the relative duration of a vowel we should compare it to the duration of other vowels in the sentence. We can do this by first calculating the average vowel duration of short and long vowels of each sentence (after the tempo normalisation has been applied). The vowels that are considered short are (@), I, A, O, Y and E[8]; the vowels that are considered long are a:, e:, o: and 2:, some vowels are of average length like i.,y. and u. [1]. For simplification purposes these vowels are considered long in this calculation. Because Ei is a compounding of two vowels it is considered long. This calculation yields average short and long vowel durations as listed in the table underneath (table 3.8).

Speaker	tempo	Average short vowel duration	Average long vowel duration
Newsreader	6,0 s.p.s.	0,084	0,139
Storyteller	4,3 s.p.s.	0,082	0,121

Table 3.8. Average short and long vowel duration after tempo normalization

Next step is to divide all normalised short and long vowel durations by the corresponding average vowel duration, resulting in relative normalised vowel durations (appendix A.4). In figure 3.6 and 3.7 the relative duration values are plotted for each vowel of the two sentences.

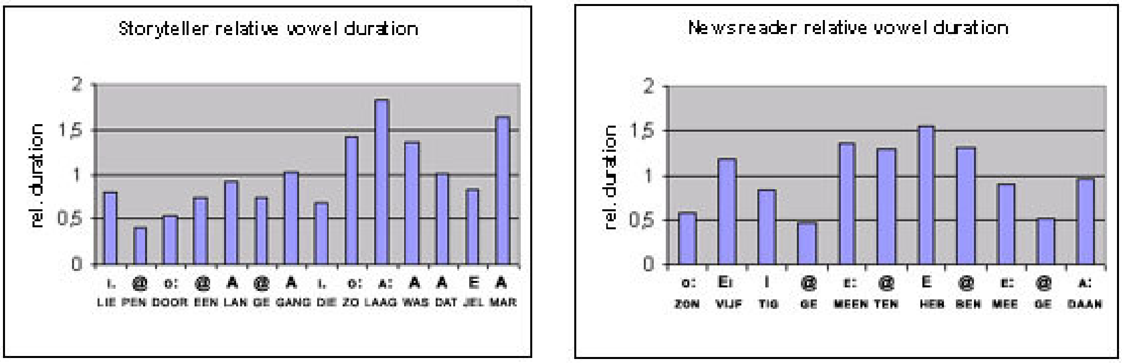


Figure 3.6 and 3.7. Relative vowel duration of newsreader and storyteller

In the diagram the relative duration axis of 1 represents the average vowel duration of the vowels to which the relative duration can be compared. For example, if a certain vowel has a duration of 1,2 this means the vowel lasts 1,2 times longer than the average duration of this kind of vowel.
Looking at the vowel durations in the diagram we see that the newsreader has a regular duration pattern which is always inside the relative duration range of [0,5 , 1,5]. The storyteller fragment shows a more increasing pattern of vowel duration, resulting in top duration values for the words “zo” and “laag” of respectively 1,4 and 1,8 times the duration of the average.
Grammatically the part of the sentence in which this duration increase occurs is in accented long vowels of the adjectival part, for example in the predicate of the sentence or in the adverb (followed by an adjective and noun) and the adjective (followed by a noun). Intuitively the duration increase is likely to occur in adjective parts of the sentence, because those words create meaning in a story and therefore can be emphasised extra.

[8] in SAMPA [1] notation for Dutch

Because we have only compared two small speech fragments there is not enough foundation to base any conclusions on, but we can say there is ground to assume duration increases in accented long vowels are based on the part of speech category of the words.

Looking at this from the perspective of the Virtual Storyteller project, the creation of a certain plot and accompanying grammatical analysis of the sentences is performed in a preceding phase of the storyteller environment. So we assume that in the generation of the plot which is written down in a certain annotation, there is the possibility to indicate that a certain accented syllable should be lengthened, based on a grammatical analysis that is performed in that phase. So for the analysis of narrative speech it is enough to include the possibility of increasing the vowel duration, resulting in a conversion rule which is applicable depending on the input test annotation.

3.5.7 Position and contour form of features in the utterance

After having observed that in general storytellers have more variation in the pitch and intensity levels than the newsreader has (§3.5.2 and §3.5.3), the next step is to find out whether these variations occur in a specific position of the utterance. It seems reasonable to accept as a hypothesis that most emphasis will be on syllables of words that have a key function in the sentence.

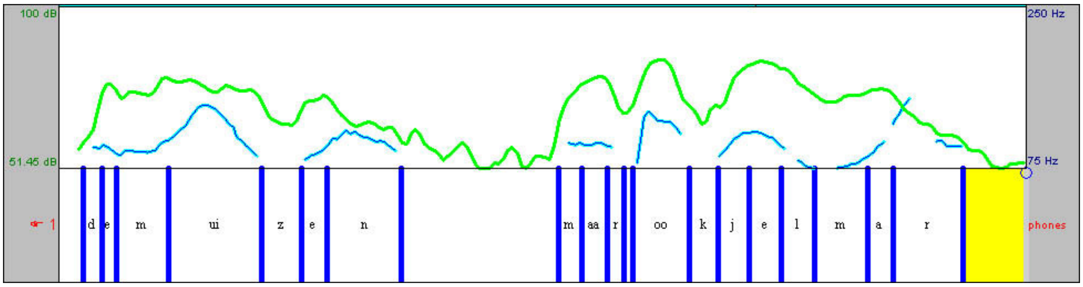


Figure 3.8. Child storyteller pitch and intensity contour

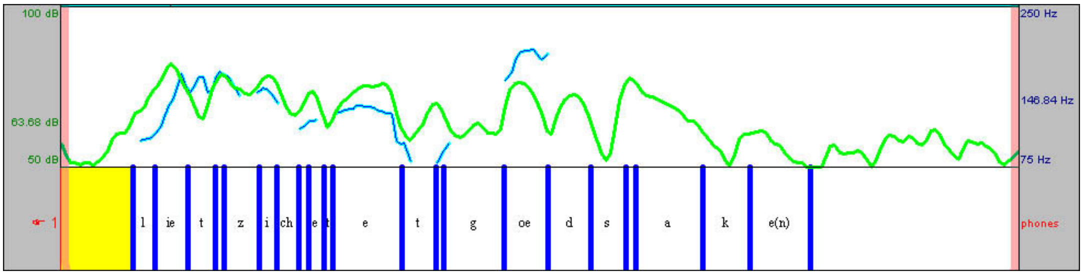


Figure 3.9. Child storyteller pitch and intensity contour

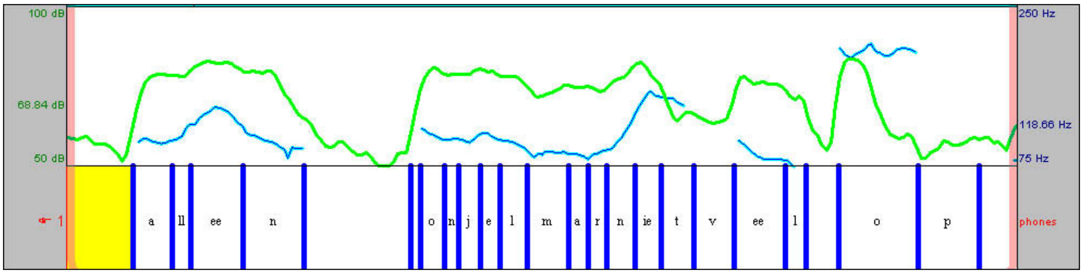


Figure 3.10. Child storyteller pitch and intensity contour

Figures 3.8, 3.9 and 3.10 show the pitch and intensity contour of the utterance ‘De muizen, maar ook Jelmar, lieten zich het eten goed smaken. Alleen kon Jelmar niet veel op.’.

Depending on the meaning the speaker wants to give to the sentence, different ways of putting the sentence accents are possible. When listening to the utterance, the speaker puts accents on word accent syllables of ‘muizen’, ‘Jelmar’, ‘goed’, ‘alleen’ and ‘op’. Looking at the figures of the contours the pitch contours have peaks on these words. The contour rises before the vowel, and during the pronunciation of the vowel it is at its maximum value, followed by a fall that continues into the beginning of the next phoneme. This is the typical high tone H*, which means there is a peak on the accented syllable [18]. The intensity contour also has peaks on accented syllables, but has a more constant course than the pitch contour does.

Comparing these contours to that of the newsreader (fig. 3.3) it is clear that the newsreader also uses H* accents, only with less variation in pitch range (the same more or less goes for the intensity contour). There seems to be an almost constant pitch and intensity rate in the newsreader’s contours, with noticeable variation in pitch on the places of the accented words of the sentence (‘vijftig’, ‘gemeenten’ ‘meegedaan’, ‘autoloze’, ‘zondag’). The pitch and intensity contour of the storyteller have much more variation, also in non-accented words, but this variation is most evident in sentence accent positions.

In general the form of a pitch contour in the sentence accent positions can be estimated by a sine or cosine function. Looking at the form of the pitch contour, in most cases there is a rise of a certain length followed by a fall of equal length (figure 3.8, “m**ui**zen” or “**jel**mar”). In the sine function this corresponds to sine values of ¼ pi to ¾ pi. Another trend that can be seen in the pitch contour is a relatively long rise followed by a short fall (figure 3.10, “**niet**”), corresponding with sine values of 0 to ¾ pi.

3.6 Summary

The position in sentences in which a difference between newsreader and storyteller is observable is in sentence accents. Based on the global analysis, the quantitative pitch analysis and the analysis of the pitch contour, we observe that the storyteller has much more variation in its pitch resulting in a larger spreading and absolute mean slope of the pitch contour. The absolute pitch rise during sentence accents is observed to be in the range of 40-90 Hz.

The intensity of a storyteller is also showing more variation than that of a newsreader. The standard deviation of the newsreader’s intensity values is 4,6 dB, against standard deviation values of 7,6 dB of the storyteller’s intensity. The global tempo of the newsreader is higher than that of the storyteller; the newsreader speaks 89% faster than the child storyteller (which maintains a tempo of 3,04 syllables per second). The inside pause of the child storyteller is 39% longer than the pause of the newsreader. This percentage is 147% for the outside pauses. On certain stressed vowels (likely depending on the part of speech of the word the vowel is in) a duration increase of about 50% is observed.

4 Analysis of tension course

4.1 Introduction

The tension course of a story is determined by the events that take place in the story. When something good or bad is happening in the plot of the story, the listener (or reader) of the story can be taken towards that event in several ways. One is that the listener is gradually taken there by making him feel that something is about to happen. The other is letting the event occur suddenly, without creating the expectation. These are two examples of tension course in a story. So the tension course is the development of tension during the story, based on the meaning of the story.

The approach that is followed in the description of the tension course analysis is the same as that of the description of the narrative speaking style. A lot of aspects of the tension course analysis setup are the same as the narrative style analysis though. The only differing aspect is the nature of the analysis material that is used. Therefore we will only describe the analysis material here (§4.2), the prosodic features that are analysed and the analysis tool are the same as used in the analysis of narrative speaking style (§3.2 and §3.4). The prosodic features that are examined are the fundamental pitch value and range, intensity, pausing and tempo of the speech signal. Two aspects of tempo will be considered: the general tempo and the vowel duration variance. Since the same prosodic features as examined in the analysis of narrative speaking style are to be examined, we will use the same tool to perform the tension course analysis, namely *Praat* [33]. This chapter is concluded with the discussion of the analysis results (§4.3).

4.2 Analysis material

For the investigation of the tension course we have selected specific fragments of child stories that represent a certain tension phenomenon. The most regularly occurring phenomenon is the climax, which is a peak or a very dramatic, thrilling moment in the story. We distinguish two types of climaxes, a sudden climax and a gradually increasing climax. The sudden climax is a climax that occurs when something unexpected occurs in the plot. The event disturbs the current quiet state of the plot, because it is this important that extra attention should be paid to the event. After the event has taken place the tension course returns to the quiet state. The second type of climax is the increasing climax. Similar to the sudden climax a special event is about to take place, but this time the listener is gradually taken towards the event. After the expectation of the listener is increased to the maximum the event is revealed. Afterwards the expectation is gone so the tension course returns to normal.

To illustrate the use of tension course better, underneath are some examples of sentences containing climaxes[9]. The words indicated in italics are positions in which a climax occurs.

[9] All sentences are taken from the fairy tale “De boze reus werd een lieve reus” and are spoken bij Frans van Dusschoten, male Dutch storyteller, “Luister sprookjes en vertellingen”, Lekturama.

“Op een dag werd de reus wakker. *Opeens* merkte hij dat hij het niet meer koud had. En *toen* hoorde hij een merel zingen.”

“Maar *toen* de kinderen hem zagen werden ze bang en holden weg.”

“Hij pakte een grote hamer *en* sloeg de muur en het hek kapot”

“Op een koude wintermorgen zat de reus naar buiten te kijken. *Opeens* zag hij in een hoek van de tuin een boom met gouden takken.”

“Zo’n mooie boom had de reus nog nooit gezien *en* onder de boom stond het kleine jongetje.”

In order to investigate the climaxes in more detail we have selected two fragments from the ‘Blauwbaard’ fairy tale[10] and a fragment from ‘Brammetje Braam’[11]. Before those fragments are analysed a global analysis of several fragments that contain a climax is performed.

4.3 Analysis results

4.3.1 Global analysis

The form of the climaxes has been described already in paragraph 4.2, now we will have a look at the acoustic realisation of the climaxes.
From global analysis of several fragments that contain a climax (fragments of paragraph 4.2) it turns out that at the moment a sudden event takes place in the story, the storyteller raises both pitch and intensity of his speech (sudden climax). In the case of an increasing climax the pitch and intensity are gradually increased while approaching the event, and gradually decreased to normal afterwards.
While the sudden climax usually occurs in one word, the increasing climax can be spread over several words. It seems like this kind of climax is sometimes spread over more than one sentence, but in fact in this case the sentences to which the climax applies are separated by a comma (which is observed by absence of final pitch lowering during the sentence and short pausing afterwards). Consider for example the following sequence of sentences:

“Stapje voor stapje kwam hij dichterbij het vage lichtschijnsel. Nog één stap te gaan en, toen kwam plotseling de wonderschone prinses te voorschijn.”

The first sentence is responsible for an increase of tension caused by its meaning, but isn’t part of the climax, because at the end of the sentence there is a decrease of pitch. The climax as we consider it is the phenomenon that is observable in the second sentence. It is strictly increasing in pitch in its first half, and strictly decreasing in its second half.

[10] “Blauwbaard”, André van de Heuvel, male Dutch storyteller, “Sprookjes van moeder de gans”, WSP select

[11] “Brammetje Braam”, male Dutch storyteller, “Luister sprookjes en vertellingen”, Lekturama

4.3.2 Sudden climax

In the first fragment a girl enters a secret room and she finds something horrible. The storyteller speaks without tension when she enters the room, but then the storyteller suddenly raises his voice and tells about the horrible discovery. The voice and pitch rise is clearly visible in figure 4.1 which contains the end of the pitch (dark grey) and intensity (light grey) contour of the utterance ‘ze moest even wennen aan de duisternis en toen..’, followed by the revelation of the secret.

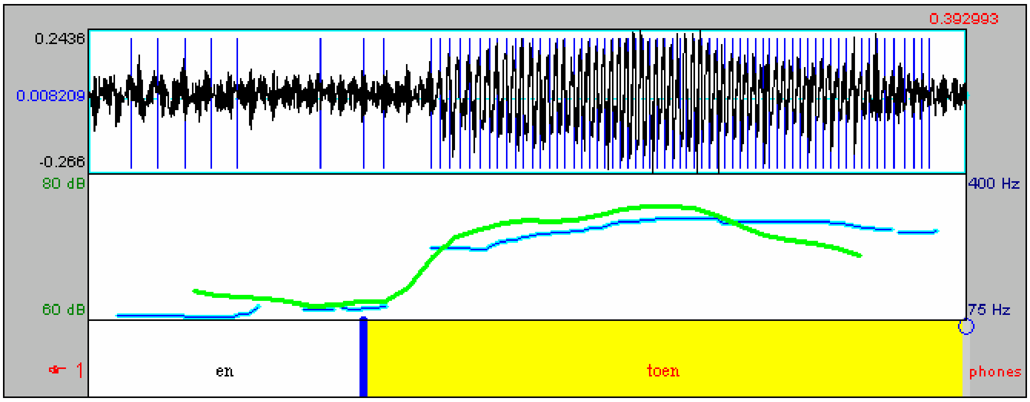


Figure 4.1. Pitch and intensity contour for a climax utterance

Taking a closer look at the contours, the most left line of the pitch contour represents an average pitch value of 85,2 Hz. The following line represents an average pitch of 100,7 Hz and after the rise the line has an average of 278,6 Hz. So there is already a relatively small rise in the voice during the word ‘en’, after the rise of the contour the pitch is increased by approximately 175%.
The average intensity before the rise is 62,7 dB, afterwards it is 73,1 dB, an increase of around 17%, which in fact means that the second part of the utterance is over ten times more intense than the first part.
One last observation can be made regarding the duration of the utterance. The pronunciation of ‘toen’ by a storyteller in a normal sentence takes 0,17 seconds on average; the duration of the vowel is 0,08 seconds. The ‘toen’ in the climax takes 0,26 seconds; the vowel is 0,19 seconds long. So it is clear that the ‘toen’ utterance in the climax is stretched, most of the delay is in the vowel.
The second fragment taken from ‘Brammetje Braam’ (‘ze holden naar binnen en toen.. toen rolden ze bijna om van het lachen’) contains the same kind of event, somebody enters a room and a sudden event takes place (figure 4.2).

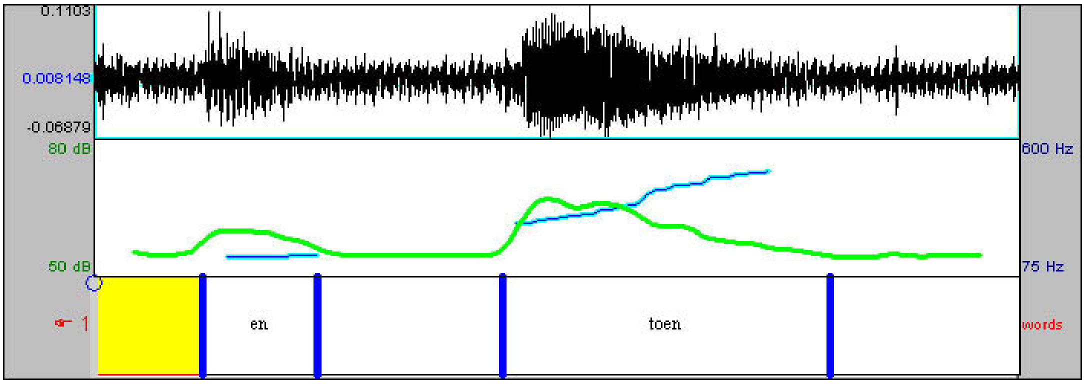


Figure 4.2. Pitch and intensity contour for a climax utterance

The pitch during ‘en’ is constant and about 150 Hz. The word ‘toen’ starts at a pitch of 277 Hz and ends at 469 Hz. So during the pronunciation of ‘toen’ there is a gradual pitch rise. The intensity during ‘en’ is 58,8 dB, rising to 65,0 dB during ‘toen’, afterwards falling back to about 58dB.

The duration of ‘toen’ is 0,35 seconds, with a duration of 0,22 seconds of the vowel. So with regard to the duration, the same conclusion can be drawn as in the previous fragment, namely that the utterance is stretched in the vowel.

Comparing the two fragments, we see that there’s a difference in the pitch contour directly after the pitch rise, in the first the pitch stays at the same level, in the second it is raised even more (H* H L% and H* H H% respectively [18]). This difference can be explained by the meaning of the words of the climax, the second fragment is a case of question intonation and therefore has an end increase.

4.3.3 Increasing climax

The third fragment (‘Hij deed de deur open en... daar lag de slapende prinses.’, figure 4.3) is different from the first two in the fact that the event that takes place in the story is already expected by the listener. The storyteller raises his voice more and more and after a pause that is used to increase the tension he reveals the secret. In contrast with the first fragment, in this fragment besides pitch and intensity variation there is a clear variation in tempo.

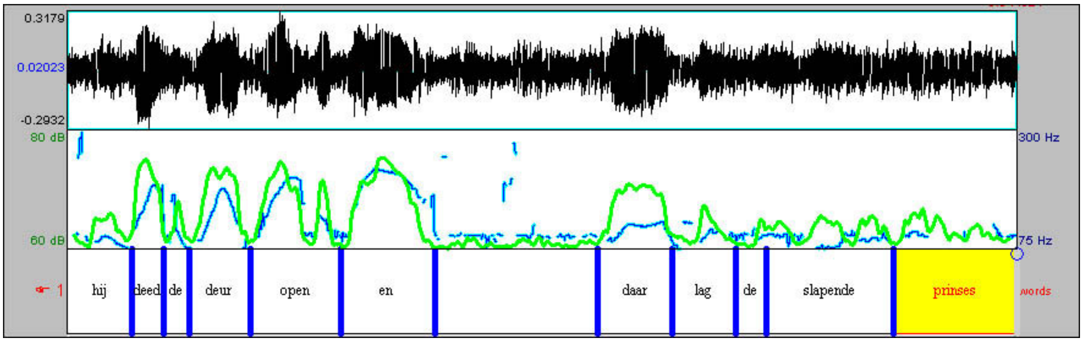


Figure 4.3. Pitch and intensity contour for a climax utterance

From the beginning of the fragment, the tension is increased until the pause after ‘en’, here the climax reaches its top, afterwards the tension returns to normal.

In the pitch contour in the figure the peak value of ‘deed’, ‘deur’, ‘open’ and ‘en’ on the accented syllables of the words are respectively 196,4 Hz, 188,5 Hz, 209,9 Hz and 225,8 Hz. So there is a small decrease in pitch followed by a substantial increase. The second part of the utterance has a lower pitch average, the peak on ‘daar’ has a value of 125,8 Hz, afterwards decreasing to an average of 98 Hz in the rest of the utterance.

The peak intensity value is about 73 dB for the words ‘deed’, ‘deur’, ‘open’ and ‘en’; a clear pattern of rise and fall can be distinguished on all words in the first part of the utterance. In the second part only ‘daar’ is spoken more intensely (69,4 dB), afterwards the signal is weakened to about 63 dB.

As said before, the speaker uses tempo variation and pause during the utterance. The pause is in the middle of the utterance and lasts for 1,04 seconds. The average tempo of the utterance is 2,99 syllables per second (after removing the pause), considerably slower than the average (4,6 syll. per sec. for this storyteller).

In the section about vowel duration we have determined average vowel length for both short and long vowels for a storyteller. Assuming the speaker of this fragment also has the tendency to make both short and long vowels have equal length, we can compare the length of vowels in the fragment to the average values (approximately 0,08 sec. for short and approximately 0,13 sec. for long vowels). The length of the vowels is listed in table 4.1, long vowels are indicated with grey shading.

syllables	Hij	deed	de	deur	o	pen	en
vowel	Ei	e:	@	2:	o:	@	E
duration	0,107	0,162	0,104	0,249	0,322	0,048	0,197

syllables	daar	lag	de	sla	pen	de	prin	ses
vowel	a:	A	@	a:	@	@	I	E
duration	0,339	0,070	0,059	0,262	0,068	0,061	0,074	0,111

Table 4.1. Duration of vowels in ‘Hij deed de deur open en daar lag de slapende prinses.’

Almost all long vowels have an above average length, besides it is remarkable that the consonants at the end of the two parts of the utterance last relatively long compared to other consonants in the utterance. The ‘n’ following the ‘e’ lasts 0,304 seconds, the ‘s’ terminating the utterance has duration of 0,323 seconds.

To visualise the deviation with regard to the average value, the relative duration has been determined. This is done by dividing the duration values by its corresponding average for short or long vowel. Figure 4.4 shows the relative durations of the first part of the fragment.

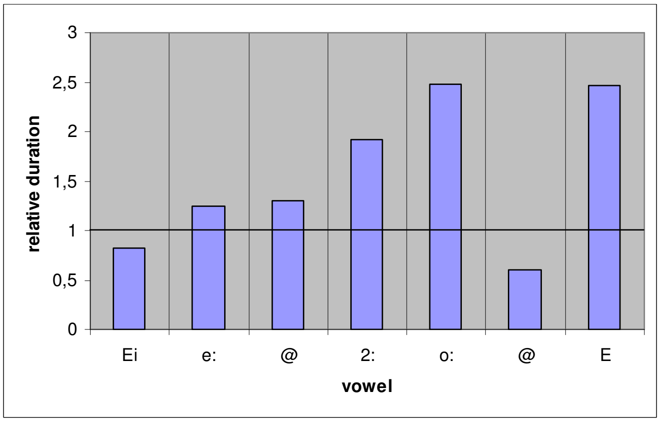


Figure 4.4. Relative duration of vowels in 'Hij deed de deur open en..'

The diagram clearly shows the increase in duration of the vowels during the pronunciation of the utterance. It is remarkable that not only words that carry sentence accent are stretched out, but almost all words in the utterance have longer vowel durations on stressed syllables (for example the word 'en' before the pause).
The 'e' vowel of 'open' seems to be conflicting with the expectation based on the other vowel durations, because it lasts relatively short with respect to its neighbours. If we take another look at the utterance it turns out that all words in the first part are one-syllable words except 'open', so all vowels have word stress except the '@' concerned (besides schwa's are never stressed).
Figure 4.5 shows the relative vowel durations of the second part of the fragment.

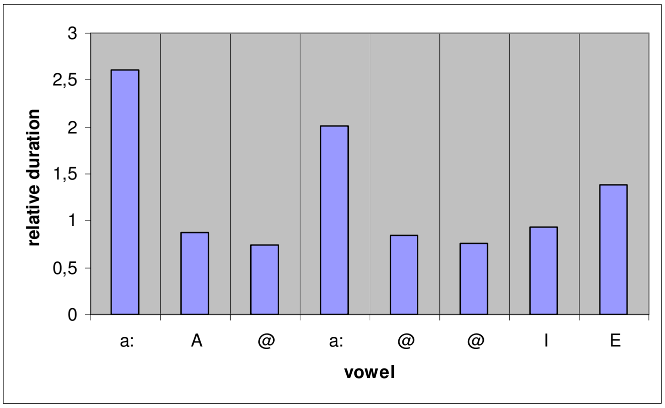


Figure 4.5. Relative duration of vowels in '.. daar lag de slapende prinses'

During the revelation of the climax the duration of vowels returns to its normal level. The diagram shows that the stressed vowels of words that carry sentence accent ('daar', 'slapende') are still relatively long. The other vowels are of normal length.

4.4 Summary

The following table (table 4.2) summarises the observed properties of a sudden climax:

Property	**Value**
Pitch	start: rise of 80 - 120 Hz, afterwards: rise 200Hz OR stay at same level
Intensity	start: rise 6 – 10 dB, afterwards: gradually decrease OR stay at same level
Duration	accented vowel after start: increase of 156% in vowel

Table 4.2. Summary of sudden climax properties

The table below (table 4.3) summarises the observed properties of an increasing climax:

Property	**Value**	

Pitch	**1st part (start until top):**
	
	2nd part (top until end):
		start at average + 100Hz
		gradual increase to +130Hz
		start at + 25Hz
		decrease to average

Intensity	**1st part:** **2nd part:**	+10 dB +6 dB on first word, rest at average intensity
Duration accented vowels	**1st part:** **2nd part:**	start at normal length gradual increase to 1,5-2x normal length decrease from 1.5-2x normal length, only on vowels with sentence accent
Pause	**between 1st and 2nd part:**	1,04 sec.

Table 4.3. Summary of increasing climax properties

5 Conversion from neutral to narrative speech

5.1 Introduction

In this section we formulate concrete rules which can be used to transform a neutrally spoken sound fragment into a fragment spoken by a storyteller. These rules are based on the outcome of the analysis of narrative speech that was conducted in the previous chapters (chapter 3 and 4) and are intended to be used in the implementation phase of the project (chapter 9). In the implementation the rules will be used to automatically create a narrative pronunciation of a certain input text. Before the rules will be used in the implementation we will first evaluate them in the next phase (chapter 6, 7 and 8) to find out whether they increase the quality of storytelling.

This chapter starts the description of the rules that are formulated for acoustic features pitch, intensity, tempo, pausing and duration of vowels of narrative speaking style (§5.2). The two paragraphs that follow contain rules for the sudden climax (§5.3) and increasing climax (§0). The chapter is concluded (§5.5) with a choice for one of the text-to-speech engines mentioned in the text-to-speech section (§2.6). The engine that is chosen will be used during the rest of the project. This choice is based on the outcomes of the analysis phase.

The starting point in the construction of the conversion rules is that we have to specify what kind of data they will apply to. As said before, the rules will be used to convert neutral speech into narrative speech, but just applying the rules to a waveform speech signal will not work, because each rule applies to a certain acoustic aspect of the speech signal. The speech input format that is used for the rules is the prosodic information of speech, consisting of pitch, intensity and phoneme duration values of speech.

For the acoustic features pitch and intensity that are processed by the rules we assume the feature data is supplied as a series of paired time-value data. This is a fundamental representation of the pitch and intensity data which allows us to formulate the rules is such way that they can be used in both the following evaluation phase as the implementation phase of the project.

So the feature data `y` is a function of time `t`:

```
y(t), t = 0 .. b

where

t      time in seconds
b      end of speech fragment
y(t)   acoustic feature value y as function of t
```

But for the application of rules we need more information. From the analysis it turns out that the prosodic functions that are present in narrative speech influence the acoustic features locally. This means for example that climaxes must be applied to a certain part of the speech only, and that narrative style is observable in sentence accents (§3.5.7). So before the rules can be applied there should be knowledge about the positions in which they should be applied (which is supplied by an annotator or by automatically generated output of a preceding stage of the project).

All the rules we use are of elementary nature, meaning they manipulate only one specific time domain (or unit, which can for example be a syllable) at a time, so they can be applied to any series of values of a certain acoustic feature as long as the start and end position of the adaptation are known. So each time a certain conversion rule for one of the acoustic features is applied, the domain of the adaptation must be supplied:

$[t_1,t_2]$

where
t_1 start time of manipulation
t_2 end time of manipulation

The determination of the time domains to which the manipulations must be applied is a process that takes place in a stage before the application of the rules. This is the responsibility of the annotator, who will annotate sentence accents and climaxes in the input text.

Some of the rules that are given below contain constants for which only a global value range is given. This is because the analysis showed a lot of variation among these values in their realisation in speech. If possible, the exact values of these constants will be determined in the constant evaluation that follows the current project phase (chapter 7). At this point it can't be said that for these constants there exist exact best values which always hold. Maybe we will have to use a stochastic model which produces values between certain limits for some constants, because the variation in their values is also responsible for the variation in speech.

When manipulating speech fragments it is important that in case of a pitch or intensity manipulation we don't base the manipulation on relative results but on absolute results of the analysis. So for example, if from the analysis it turns out that a certain adaptation requires that a storyteller increases his pitch value by 80Hz with respect to his average pitch value of 100Hz, we must not make a rule which increases the average pitch in question by 80%, but the rule should realise an absolute pitch increase of 80 Hz. This is because each speaker has a personal intrinsic base level of pitch and intensity, so relative manipulations would yield larger absolute adjustments for speakers with a higher base level, which is not desirable because then a too large adjustment is applied.

5.2 Narrative speaking style

5.2.1 Pitch

In the transformation of the pitch contour accented syllables of key words in the sentence should be manipulated (§3.5.1 and §3.5.7). This is realised by multiplying all original pitch values by a certain time dependent factor. The value of this factor depends on the form of the pitch contour we want to create. From the feature position and contour analysis (§3.5.7) it turns out that the pitch contour during a syllable that has sentence accent can be described by using a sine or cosine function, since the pitch rises to a certain peak and afterwards decreases.

In order to let a certain syllable take this form, we multiply all values inside the syllable time domain $[t_1,t_2]$ by a factor based on the sine waveform (the factor value is explained below). The adaptation process of the pitch values is based on the fact that the pitch contour of a certain signal is represented by a pitch value `y` (Hz.) for each time unit `t`. So we can process all time frames one by one and then modify the corresponding pitch values.

A single syllable pitch manipulation for a syllable in domain $[t_1,t_2]$ is performed by the following formula:

$$y'(t) = \begin{cases} y(t).(1+(\sin((((t-t_1)/(t_2-t_1))m_2 pi)+m_1 pi)/n)) & ,if \quad t \in [t_1,t_2] \\ y(t) & ,else \end{cases}$$

The following variables and constants are used:

`y`	acoustic feature value y as function of t
`y'`	manipulated pitch values
m_1	constant determining starting point of sine waveform that should be used
m_2	constant determining fraction of sine waveform that should be used
`n`	constant determining the degree of adaptation

By applying this formula for all values of `t`, the pitch values are manipulated inside the syllable domain and the result pitch values are in `y'`.

From the analysis we know that most pitch movements on accented syllables can be approximated by a hill form sine function (or a part of it). So in the formula for pitch manipulation we have used a sine function to realise this pitch contour. Looking at a plot of sine values from *0* to *2pi* in figure 5.1 we can see that the pitch contours of a storyteller on accented syllables can best be approximated by the part of sine plot from *0* to *pi*. In the formula constant m_1 and m_2 determine the form of the contour by taking on values that represent the fraction of the sine plot that is desired. From the analysis it turns out that most accented syllables are described by a sine form of ¼ pi to ¾ pi or by a sine form of 0 to ¾ pi. This means that for each word that has a sentence accent and therefore is manipulated, one of the two pitch contour forms has to be chosen.

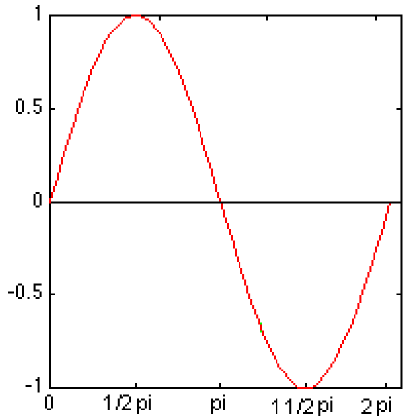


Figure 5.1. Sine function

In the formula, we want the sine function to return values between 0 and 1. This means the input values of the sine function must be normalised between 0 and `pi`. We normalise the time frames `t` by subtracting the start time t_1 from any `t` value and then divide it by the total duration of the domain (t_2-t_1), which ensures that the outcome is always a value between 0 and 1.

Furthermore, in the formula the new pitch value is calculated by increasing the old pitch value by the outcome of the sine function divided by a constant *n*. This constant determines the influence of the outcome of the sine function, so in fact it is used to determine the magnitude of the pitch adjustment.

If for example we want to realise a pitch increase of maximally 80Hz. and the speaker has an average base pitch of 100 Hz., we can solve the following equation to determine the value of *n*:

```
180 = 100*(1 + sin(½ pi) / n)
```

It is easy to see that in general we can rewrite this to:

```
n = avg_pitch / desired_max_pitch_increase
```

As already explained in the introduction, the exact values of some constants are still to be determined in the constant evaluation (chapter 7). Therefore we will now only give the limits between which the constants must lie (table 5.1):

Constant	Description	Range
m_1	starting point of the sine curve to use	0 or 0,25
m_2	fraction of the sine curve to use	0,75 or 0,5
`desired_max_pitch_increase`	The maximum increase with respect to the average pitch	40 – 90 Hz

Table 5.1. Pitch constants

5.2.2 Intensity

From the general intensity analysis results (§3.5.3) and the study of position and contour form of the intensity (§3.5.7) we know that accented syllables and their surrounding phonemes have relatively high intensity and that the standard deviation of the storyteller's intensity has an average of about 7 dB. The syllables that have sentence accent have an increase in intensity value, but we didn't exactly determine the average value of this increase in the analysis. Therefore as a maximum value for the intensity increase we will use the value of the standard deviation. A standard deviation of 7 dB means that 95% of the intensity values are inside a 14dB range around the mean intensity value (7 dB above and 7 dB below the mean). So if we use an increase of 7 dB in our rule we know that the output intensity is inside or near the range of the majority of intensity values, which is a safe value to use.

Compared to the pitch contour the intensity contour shows a more constant course, so it isn't necessary to use a sine form contour here, a simple addition of the intensity values will suffice. Another reason for not using a sine contour is that the degree of intensity adaptation is small compared to the adaptation degree of the pitch. Because adaptation is small the contour that is already present is not altered rigorously so the contour course that was already present is still there after the adaptation.

So we can increase the intensity values in a certain domain by a certain value to get the desired result. In order to include surrounding phonemes, we will take the same domain as is used for pitch manipulation and use a constant k with which we will augment this domain.

A single syllable intensity manipulation for a syllable in domain $[t_1,t_2]$ is performed by the following formula:

$$y'(t) = \begin{cases} y(t) + c & ,if \quad t \in [t_1 - k, t_2 + k] \\ y(t) & ,else \end{cases}$$

with

y'	manipulated intensity values
k	constant determining the increase of the domain
c	constant by which intensity value is increased

The following table (table 5.2) lists the range or value of the constants that have to be used:

Constant	Description	Range
c	intensity value increase	4 – 7 dB
k	time domain increase	0-0,2 sec

Table 5.2. Intensity constants

5.2.3 Tempo

The adjustments that have to be made to the tempo are quite simple. We have to slow down the general tempo of the signal by a certain amount.

The adjustments of the rest of the prosodic features, including tempo, are conversion rules that we will not formulate in the way we did as the rules for pitch and intensity. The reason for this is that the values of the acoustic features in question are not expressed as a function of time. For example, the tempo of a speech signal in this case isn’t expressed as a function of time, because it is the general tempo of the speech which is constant.

We assume that the tempo of the neutral input speech is expressed in syllables per second, so here we will only provide the general tempo that should be prolonged in the same quantity. From the analysis (§3.5.4) we know that the child storyteller has an average tempo of 3,0 syllables per second, the adult storyteller has an average tempo of 3,6 syllables per second. So the tempo of the output speech should be between 3,0 and 3,6 syllables per second.

5.2.4 Pausing

By slowing down the tempo we have already lengthened the pauses of the output signal, but from the analysis (§3.5.5) it turns out this is not enough, because the difference of the outside pauses between the newsreader and storyteller is larger than the difference in tempo. So we must also make sure that an outside pause of about 1,3 seconds is present in the output signal. Pauses inside a sentence should take about 0,4 seconds (based on both the adult as the child storyteller).

5.2.5 Duration of vowels

Based on the findings in the analysis (§3.5.6) in certain adjectival parts of the sentence a duration increase of an accented vowel is required (the syllables of parts for which this is the case are annotated as such). If this is the case the duration of the concerning vowel may increased by a factor 1,5.

5.3 Sudden climax

In case of the realisation of a sudden climax the pitch, intensity and duration of a part of a certain utterance have to be manipulated. In the analysis (§4.3.2) it turns out that not all acoustic properties show the same behaviour all the time, but that there is some variation. For this reason we will need to include these variations in our evaluation and see whether there is a 'best' choice or whether we have to use a more probabilistic approach, which means letting different variations occur within a certain range.
In the description of the climax we use a time domain to indicate the climax, this time domain is to be indicated in the annotation:

```
[t1,t2]
where
t1     start time of climax
t2     end time of climax
```

In this time domain the start of a climax is defined as the moment that an unexpected event occurs in the plot and as a result a sudden change finds place in the acoustic features pitch, intensity and duration. After the climax ends the acoustic features return to normal.

5.3.1 Pitch

In the analysis we observe a pitch rise of 80-120 Hz at time t_1. From t_1 to t_2 the pitch can be increased by another 200Hz or stay at the same level. These manipulations can all be done by simply shifting the pitch contour inside a certain time domain by these values (no sine formula is needed). The following formula can be used to manipulate the pitch of the sudden climax (no increase from t_1 to t_2 is performed):

$$y'(t) = \begin{cases} y(t) + c & ,if \quad t \in [t_1, t_2] \\ y(t) & ,else \end{cases}$$

with

y'   manipulated pitch values

c   constant by which pitch value is increased

If we want to gradually increase the pitch afterwards during [t_1 , t_2] the following formula must be used:

$$y'(t) = \begin{cases} y(t) + c + d(t - t_1)/(t_2 - t_1) & ,if \quad t \in [t_1, t_2] \\ y(t) & ,else \end{cases}$$

with

y'	manipulated pitch values
c	constant by which pitch value is instantly increased
d	constant by which pitch value is gradually increased

Here constant d is multiplied by the quotient of $t-t_1$ and the length of the climax t_2-t_1, resulting in a gradual increase from 0 to d spread over the length of the climax.

5.3.2 Intensity

At t_1 the intensity is increased. This increase can be done by shifting the intensity contour by 6-10 dB. From t_1 to t_2 one option is to let the contour gradually return to its old intensity, the other is to stay constant. The formula that was used for the pitch adaptation can be used for the intensity as well, because the manipulation and the domain of both features are the same. The decrease of intensity from t_1 to t_2 can be formulated analogously to that of the pitch increase, but no extra constant for the decrease is needed because want to decrease the value of c (because of their similarity to the pitch formulas the intensity formulas will not be given here). The following variables and constants are used instead for the intensity manipulation:

y'	manipulated intensity values
c	constant by which intensity value is increased

5.3.3 Duration

The exact durations we measured in the analysis were based on the analysis of the utterance ‘en toen’. Since the duration increase should be applicable independently of the words it applies to, we can not use the absolute duration increase values that were found for the ‘en toen’ utterance. Because we want to apply the rule to different utterances we will derive a rule based on relative duration increase.

It turns out that the total duration of ‘toen’ is increased by 79%, and that the duration of the vowel is increased by 156%. From the results it also turns out that the duration of consonants ‘t’ and ‘n’ is not increased (0,09s normal, 0,1s climax), so that all increase is in the duration of the vowel.

5.3.4 Summary

The following table (table 5.3) summarises the sudden climax properties:

Property	Value
Pitch	t_1: rise of 80 - 120 Hz, $[t_1, t_2]$: rise 200Hz OR stay at same level
Intensity	t_1: rise 6 – 10 dB, $[t_1, t_2]$: gradually decrease OR stay at same level
Duration	accented vowel after start: increase of 156% in vowel

Table 5.3. Sudden climax constants

5.4 Increasing climax

In the description of the climax we use a time domain to indicate the climax:

```
[t1,t2] and [t2,t3]
where
t1      start time of climax
t2      time of climax top
t3      end time of climax
```

In this time domain the start of a climax is defined as the moment that the tension starts to gradually increase, until it reaches its peak tension in the climax top. After the climax top the tension is still present but gradually decreases until the normal state is reached again.

5.4.1 Pitch

At time t_1 there is an initial pitch increase which is about 100Hz higher than the average pitch (§4.3.3). From t_1 to t_2 this pitch is gradually increased by about 30Hz. The initial pitch at time t_2 is 25 Hz higher than the average pitch, decreasing to the average at t_3.
The difference with the previous climax (sudden climax) is that a simple pitch increase formula for domain $[t_1, t_2]$ and $[t_2, t_3]$ can't be given. The problem is that the sudden climax applies to one or two words only, but the increasing climax can apply to a complete sentence. If we would create a similar formula for the increasing climax as we did for the sudden climax, for the domain $[t_1, t_2]$ this would result in a formula that increases the pitch of all words in $[t_1, t_2]$. This is not correct because the pitch manipulations of the increasing climax should only be applied to syllables that have word accent. Also the form of the pitch increase should be based on the sine function so it can not be realised by a simple pitch value addition.
In order to do the correct manipulation, we should manipulate each individual accented syllable (§4.3.3) of each word in $[t_1, t_2]$. This means that inside domain $[t_1, t_2]$ for each accented syllable we should find out what its start and end time are, giving us syllable domain $[s_1, s_2]$ for each single syllable. Based on this knowledge we can calculate the desired pitch increase for that syllable based on the position the syllable has relative to $[t_1, t_2]$.

As said before, the pitch during $[t_1,t_2]$ should rise 30 Hz. In order to find out the gradual increasing pitch value with which a certain syllable starting at time s_1 ($t_1<s_1<t_2$) should be increased, we can solve the following equation for each syllable[12]:

$$\text{desired\_max\_pitch\_increase} = 30(s_1-t_1)/(t_2-t_1)$$

Next step is to apply a pitch manipulation which has the same form as that used for narrative style (§5.2.1):

$$y'(t) = \begin{cases} y(t).(1+(\sin((((t-s_1)/(s_2-s_1))m_2 pi)+m_1 pi)/n)) & ,if \quad t \in [s_1,s_2] \\ y(t) & ,else \end{cases}$$

with

y' manipulated pitch values

m_1 constant determining starting point of sine waveform that should be used

m_2 constant determining fraction of sine waveform that should be used

n constant determining the degree of adaptation

Constants m_1 and m_2 can be based on the values given in paragraph 5.2.1, we can calculate the value of n as follows:

$$n = \text{avg\_pitch} \;/\; \text{desired\_max\_pitch\_increase}$$

For the syllables in time domain $[t_2,t_3]$ an analogous approach can be used which will not be given here.

5.4.2 Intensity

Because no gradual increase or decrease of intensity is needed, the manipulation of intensity is relatively simple. All syllables $[s_1,s_2]$ in $[t_1,t_2]$ should have their intensity values increased by 10 dB, there is a 6 dB intensity increase on the first accented syllable $[s_1,s_2]$ in $[t_2,t_3]$. The rest of $[t_2,t_3]$ is spoken with average intensity and isn't manipulated. The formula used to apply to the syllables is the following:

$$y'(t) = \begin{cases} y(t)+c & ,if \quad t \in [s_1,s_2] \\ y(t) & ,else \end{cases}$$

with

y' manipulated intensity values

c constant by which intensity value is increased

[12] We base the calculation of the pitch increase of the syllable on the start time of the syllable. We could as well have to chosen to use the time at the middle of the syllable ((s_1+s_2)/2). These differ so little that the difference isn't audible.

5.4.3 Duration of vowels

We will again only give a description of the duration increase. The duration of accented vowels is gradually increased to about 1,5 - 2 times the normal length during $[t_1,t_2]$. During $[t_2,t_3]$ only the accented vowels of words that have sentence accent are lengthened (decreasing from increased duration to normal length).

5.4.4 Pausing

At time t_2 a pause of about 1,04 second must be inserted.

5.4.5 Summary

The following table (table 5.4) lists a summary of the properties of the increasing climax:

Property	Value	

Pitch	**1st part ($[t_1,t_2]$):**
	
	2nd part ($[t_2,t_3]$):
		start at average + 100Hz
		gradual increase to +130Hz
		start at + 25Hz
		decrease to average

Intensity	**1st part:** **2nd part:**	+10 dB +6 dB on first word, rest at average intensity
Duration accented vowels	**1st part:** **2nd part:**	start at normal length gradual increase to 1,5-2x normal length decrease from 1.5-2x normal length, only on vowels with sentence accent
Pause	**between 1st and 2nd part:**	1,04 sec.

Table 5.4. Increasing climax constants

5.5 Text-to-speech engine selection

The most important aspect in the selection of the appropriate text-to-speech engine for our purposes is the requirement that the engine we use offers sufficient flexibility for the adaptation of acoustic properties of synthetic speech. The flexibility of an engine is considered sufficient if we have control over the properties that in the analysis were found to be important and if we have a suitable way of controlling them.

During the process of creating narrative speech in our implementation (chapter 9) the following steps are taken. First a preliminary pronunciation of the input text is obtained by having the text-to-speech engine synthesise it, resulting in a neutrally spoken version of the text. The essence of this step is that a pronunciation of the text containing its regular prosodic information is obtained[13]. The next step is to apply our conversion rules to this preliminary output and

[13] This is a requirement of our conversion rules; they should be applied to a speech signal that already contains ‘neutral’ prosodic information. So the output of the text-to-speech engine is considered neutral.

resynthesize the result. This manipulation step brings us to a part of the process that is important for the choice of our text-to-speech engine. In order to perform manipulations on the acoustic properties of the preliminary output, the engine should be able to store the prosodic information of the preliminary output in an editable format. In this way we can use the prosodic information as input for our conversion rules, and after applying them we can output the manipulated prosodic information using the same format. Of course the text-to-speech engine should then be able to resynthesize the manipulated prosodic information resulting in a narrative pronunciation of the text.

Looking back at the outcome of the analysis the acoustic properties that are important in the generation of narrative speech are pitch, intensity, duration, tempo and pausing. These are the properties that should be adaptable, and the adaptation of these should be possible on an absolute value level. So it is not enough to specify the acoustic properties by means of a descriptive term, but the conversion rules demand that we can for example process pitch values in Hertz.

Considering the first text-to-speech engine Nextens (§2.6.3), an advantage is that Nextens already supports the use of XML annotated text files for input. The SABLE markup language that is used offers the possibility to specify pitch, intensity, tempo and pausing (in absolute values), but lacks the possibility to specify durations of phonemes. Another problem here is that the SABLE markup language is used as a way to specify the input of the speech engine, but the output of Nextens can not be returned in a prosodic format (a list of phonemes with their accompanying pitch and duration values). Since Nextens doesn't return a preliminary pronunciation of the input text in a prosodic format, we can't apply our conversion rules to anything, since we don't have the neutral prosodic information that is required as input for these rules. This makes the use of Nextens impossible[14].

The other text-to-speech engine that is to be considered is Fluency TTS. Contrary to Nextens, Fluency doesn't support any kind of XML markup language for speech synthesis. But on the other hand the engine is able to return a prosodic representation of the speech output. This representation is a list of phonemes followed by their absolute duration and pitch values, and also includes the specification of pauses. The tempo of the speaker can also be controlled, but only globally. The only acoustic feature that isn't included in the prosodic representation is the intensity of the speech. Fluency can use the prosodic format as an input for synthesis.

Since Fluency offers an employable format of storing prosodic information, which furthermore can be used for resynthesis input after some manipulations have been performed, this engine meets our requirements. The only disadvantage of using Fluency is that there's no control of the intensity of the speech, so in our implementation we will not be able to include this acoustic aspect. So for the further course of the project Fluency will be used as text-to-speech engine.

[14] Because of the open source nature of Nextens it is possible though to make changes to Nextens itself making it meet our requirements. This would be an option if no other text-to-speech engine meets our requirements, but starting point is that we don't want to modify the text-to-speech engine itself.

6 General experimental setup

6.1 Introduction

This chapter describes the setup of the evaluations that are performed during the project. During the project three evaluations take place, each with a different purpose.

The first evaluation is the *constant evaluation,* it is intended to evaluate the constant ranges that were found in the conversion rules. As can be seen in the previous chapter in which the conversion rules are formulated, for some constants of acoustic properties no absolute values have been determined yet, but a value range is given for the constant. This is because from the analysis of narrative speech it turns out that there's variation among the concerning constant's values. The constant evaluation is used to determine the best value for each constant by generating speech fragments based on the conversion rules, but varying within the range of the constants. These fragments are then judged by participants, with the purpose of finding a best value for the constant.

The second evaluation that takes place is the *conversion rule evaluation*. After the best constant values have been determined in the previous constant evaluation, a deterministic set of conversion rules is obtained which can be applied to any speech signal. The question is whether the application of the rules really makes the speech sound as spoken by a storyteller. In the conversion rule evaluation we will test this by creating narrative speech fragments (using *Praat* [33]) based on the conversion rules and have these fragments evaluated by participants. Three aspects of each speech fragment are judged: the general naturalness, the quality of storytelling and how much tension is conveyed in the speech.

After the conversion rule evaluation is finished the implementation phase will start. In the implementation phase a module will be created that implements the conversion rules which can be used to automatically create narrative speech based on text input. More details about the implementation can be found in the corresponding chapter (chapter 9). After the module is built we want to evaluate it by creating narrative speech fragments and have them evaluated by participants. This *implementation evaluation* is the last evaluation that is performed; the difference with the conversion rule evaluation is that the fragments are now created automatically by our module and that the number of participants that is involved in the evaluation is larger. The judgement criteria that are used for this evaluation are the same as those of the conversion rule evaluation.

Now that we know which evaluations will take place we will describe their common characteristics in the general evaluation plan (§6.2). Here we have a look at the general setup, evaluation method, equipment, participants, stimuli nature and the process of stimuli creation of the evaluations[15]. This is followed by an explanation of the manipulation of the fragments using *Praat* (§6.3) and the software environment that was designed to support the evaluation (§6.4).

[15] Exception here is that the process of stimuli creation of the implementation evaluation differs from the other two evaluations, so this characteristic is dissimilar for the three evaluations.

After these processes have been described, the actual constant evaluation and conversion rule evaluation will be described in the following two chapters, including the results and conclusions for both phases (chapter 7 and 8).

6.2 General evaluation plan

6.2.1 General setup

In this section common characteristics of the evaluations are discussed. Each evaluation consists of an experiment. After the experiment has taken place the results are processed and discussed.

The experiment is performed by a participant using a computer. The computer is running a program that plays several speech fragments. Each fragment is accompanied by a question, which the participant can answer by selecting the desired answer using the mouse. The entire evaluation environment is in the Dutch language.

Before the evaluation takes place it is important that the method of working during the experiment is explained clearly to the participant. It is also important that the participant understands the judgement criteria the evaluation is based on, and that questions are asked in a comprehensible way. The participant will be explained (by means of an introduction text, see appendix F) what he is supposed to do in a very neutral way, also emphasising the fact that for good results the participant should judge in an unbiased way. In order to communicate all these requirements correctly and clear to the participants, the actual experiment is preceded by an introduction phase in which the goals and requirements are explained.

During the explanation phase there should also be emphasis on the fact that the participant shouldn't focus too much on intelligibility. The TTS output isn't always of perfect intelligibility and this is a problem that can't be improved by our manipulations. To prevent intelligibility problems during the evaluation we will show the participant the text of all speech fragments. Before we start the evaluation we will play some fragments which are not to be judged, but are only intended to let the participant get used to synthetic speech.

Because of the fact that a synthetic speech utterance sounds more natural after hearing it a couple of times, we want to restrict the number of times the participant can hear a certain fragment. We will not let the participant listen to a fragment more than three times, after that he must answer the question.

We will keep track of the number of times a participant listens to a fragment before he answers a question. In this way we can see how hard it is for a participant to judge a certain stimulus.

6.2.2 Evaluation method

In the past, several methods have been used to evaluate the naturalness of synthetic speech fragments. Especially in the field of emotional speech a lot of research has been done to see whether certain emotions are recognised. Two widely used evaluation methods are *forced response test* and *free response test* [10]. The first means a test participant has to judge a speech fragment by picking one answer from a fixed list of answers which he thinks is most applicable. One derivative of this method is to provide a participant with two fragments and let him choose the best from the perspective of a certain property (for example select the most natural sounding fragment). In the free response test, the test participant is not restricted in his answers and is allowed to give unrestricted free input.

The advantage of the forced response test is that it takes less time of the test participant, but it could also exclude some important observations a participant makes but isn't able to provide. The free response test gives the user more freedom, but is more time intensive both in performing the evaluation itself and in processing of the results afterwards.

In our experiments we will use a combination of the two response methods. Because the participant has to judge entire sentences of speech, in the case of judging naturalness it is possible that the participant perceives some words of the sentence as natural, and some as not natural at all. If we ask the participant to judge the sentence as a whole in a forced response test, but also provide a voluntary possibility to explain his answers, we get the feedback in the most appropriate way. In this way the compulsory part of the response contains the necessary judgement data and the voluntary part optionally contains extra remarks about the fragment in question.

There are several ways of doing a forced response test:

- One possibility is to provide the participant with two sound fragments and let him choose the better of the two based on a certain criterion. (binary choice)
- Another way is to let the participant hear a fragment and then let him choose the most suitable answer from a list (choose best). This method is frequently used in the recognition of emotions in speech, for example in an experiment in which a participant has to pick the emotion he perceives in a speech fragment from a list of emotions [10].
- Third way is to let the participant judge a single fragment by rating it on a certain criterion scale (Likert method [11]). For example, if we want to test quality of narration of a certain utterance we can ask the participant to give a mark in the range of 1 up to 5 (1 meaning very bad and 5 meaning very good).

When choosing for a certain method we have to consider we have two main goals in our evaluation:

1. Evaluate the added value of applying the conversion rules from the perspective of narration (conversion rule evaluation and implementation evaluation)
2. Find out the best values of constants in the conversion rules (constant determination)

The fragments we will use in our evaluations are neutral fragments on the one hand, and fragments equipped with narrative style and climaxes on the other hand. If we used the binary choice method of forced response test for the first goal, this wouldn't give us the desired results. We could for example let the participant hear a neutrally spoken fragment and the same fragment including a climax, and subsequently ask the user which fragment he experiences as most tense. It is plausible that the user will pick the second one as most tense in most cases, which might falsely bring us to the conclusion we did a good job. The problem here is that although the second fragment might sound tenser, this way of evaluation still doesn't give us any measurable useful information about *how* natural the tension sounds. It could still be far from natural, which in fact is what we want to know. So this method works only if we don't need refined results. On the other hand, for the sake of the second evaluation goal we do want to use this method, supplying the participant with two fragments of which one has a different value than the other for a certain constant from the conversion rules. The participant can then choose the better of the two[16].

The choose-best method of force response testing gives us the same kind of information as the first method, so it is less eligible for the first goal. Moreover, this method doesn't fit our evaluation goals because we don't have a set of answers which one must be chosen from. The method we will use for this goal is the Likert scale judgement method. Because being more refined the method gives us the information in the degree of detail we want.

So for the realisation of the first goal, our evaluation will be based on the Likert rating method of evaluation. For the purpose of the second goal the binary choice judgement method will be used. Besides these methods we will also provide the possibility to give free response in a text box.

6.2.3 Equipment

The evaluation will be carried out on a PC workstation equipped with headphones in a quiet environment, so any undesirable disturbances are minimised. There is a personal introduction after which the participant should be able to perform the experiment on his own. The evaluation is facilitated by a web application, which will be described in more detail in section 6.4.

6.2.4 Participants

When selecting participants for the evaluation there is the possibility to select persons with or without speech synthesis experience. Both groups have the ability to judge the quality of speech fragments in the forced response test. People without experience can sometimes give comments from an unbiased point of view, which sometimes are very useful. People within the field of research can provide more specific feedback because they know what they are talking about.

A disadvantage of using two different participant groups is that the evaluation results should be treated differently because the participant groups judge the fragments from different perspectives. Because we do not want to split the participants into two groups we decided to select people without speech synthesis experience only.

The constant evaluation will be carried out by a group of about 5 persons; the conversion rule evaluation will be carried out by a group of 8 participants. The final implementation evaluation will be carried out with the largest group of participants; the group has a size of 20 participants.

6.2.5 Stimuli nature

One problem that can be present in the evaluation of storytelling quality is that it is possible that the contents of a certain fragment influence the judgement. So it is possible that the presence of certain common fairy-tale aspects (names, characters, situations and locations) in a fragment

[16] This method only works if we vary only one constant at a time. The other constants should be assigned the same value for both fragments that are evaluated, so there is no influence of other constants on the results.

make a participant rate a fragment higher with respect to storytelling than a fragment that doesn't contain these aspects. But the contrary is also possible; if these aspects are present and the participant doesn't think the narrative style manipulations add anything to the quality of storytelling that is already present, he will rate the fragment lower.

So at this point we don't know how large the influence of semantics is and it would be interesting to test this in the experiment by dividing the fragments in two groups, one fragment group semantically neutral and one containing fairy-tale aspects. Because the evaluations are solely aimed at evaluating narrative speech and we want to minimise the amount of evaluation variables in this experiment, we decided not to include this semantic division. So although it is not proven that semantics influence the results in any way, we assume this to be true and as a consequence the fragments we evaluate will be as semantically neutral as possible.

For the evaluation of climaxes it isn't possible to select completely semantically neutral fragments. Climaxes often contain words that are closely related to storytelling (for example 'plotseling', 'toen'). It would be very hard to create natural sounding climaxes without the use of these climax related words. And if we would try so, the participant might get confused because based on prosodic features he experiences a climax without observing it semantically.

6.2.6 Creation of stimuli

In this section the steps that are involved in creating the stimuli for each of the evaluations is described. In our future implementation we will use Fluency TTS (§2.6.3) for the actual synthesising of speech. Our conversion rules (chapter 5) must be applied to a speech signal that already contains basic prosodic information, which can be obtained by having the text-to-speech engine create a preliminary (or neutral) pronunciation of a given input text. The neutral pronunciation contains prosodic information based on natural language processing methods for Dutch that are implemented in Fluency. For example, a speaker normally uses an effect in his speech that is called declination [1], which is the fact that the fundamental pitch of a spoken sentence gradually decreases as a sentence comes closer to its end. This is one of the common prosodic effects that is included in the neutral Fluency pronunciation.

The goal of the *constant evaluation* is to determine the best value for each constant used in the conversion rules by generating speech fragments based on the conversion rules, but varying within the range of the constants. So the way of working is as follows: an appropriate (§6.2.5) sentence is selected which is synthesised by Fluency resulting in a neutral pronunciation of the sentence. Now two fragments can be created by applying the conversion rules to the neutral pronunciation while one constant takes on two different values in the conversion rules. Another option is that only one manipulated fragment is created by applying the conversion rules and the other fragment is kept neutral. For the manual application of the rules a speech signal processing tool ('Praat') will be used (§3.4). The two fragments that are created can be used as stimuli in the constant evaluation.

The goal of *the conversion rule evaluation* is to find out whether the application of the conversion rules adds something to the narrative quality of the synthetic speech. In order to find this out we will use a paired approach, meaning that a certain sentence is evaluated in both neutral as in narrative form. By comparing the results of both judgements we can see if there's any difference with respect to narrative quality. The steps involved in the creation of stimuli for this evaluation are as follows: an appropriate sentence is selected which is synthesised by Fluency. This is the first of the two fragments; the other is obtained by applying the conversion rules to the neutral fragment using the signal processing tool. This results in two fragments that can be used as stimuli in the evaluation.

The approach that applies to the creation of stimuli for the *implementation evaluation* is quite straightforward. After appropriate sentences for this evaluation have been selected they can be used as input of our implemented module. The module should be able to return both the neutral as the narrative pronunciation of the input, which can directly be used as stimuli in the evaluation.

6.3 Manipulation of fragments

6.3.1 Introduction

Before we go on describing the execution and results of the three evaluation phases, in the following sections we will first describe some underlying aspects of the evaluations.

All fragments that are used in the constant evaluation and the conversion rule evaluation are manipulated based on the conversion rules using the *Praat* speech signal processing tool (§3.4). Because the manipulation of so many fragments by hand can be quite time consuming, we used *Praat*'s scripting possibilities for this purpose. In this section we will describe the way the fragments were manipulated with *Praat.*

As explained in the section about the creation of stimuli for the constant evaluation and the conversion rule evaluation (§6.2.6), the first step of the manipulation process is to select an appropriate sentence and let Fluency create a neutral pronunciation of this sentence. The next step is to determine the position of the sentence accents, which can be based on previous pronunciation by a storyteller or by choosing the most intuitively natural positions. It is also possible to use the sentence accents determined by Fluency. The problem with those accents is that they aren't always determined correctly and that from a storytelling perspective they aren't always the desired accents.

After these positions have been determined, the start time, end time and length of each accented syllable must be looked up in the fragment using *Praat.* Now that we have determined the time domain for the manipulations we are ready to run the manipulation scripts.

Each operation in *Praat* that is performed can be represented by a command which can be used inside a script. The advantage of using scripts is that a series of operations can be executed at once instead of doing them manually one by one. This saves a lot of time which is profitable in the evaluation because a lot of fragments must be manipulated.

Based on the conversion rules certain manipulations must be carried out on a "neutral" Fluency fragment. This can be any of the following manipulations:

- Pitch adaptation
- Intensity adaptation
- Duration of vowels increase
- Pause insertion

In the following section the script used to realise one of these manipulations will be explained. We will only discuss the script that was used for the manipulation of the pitch, the other scripts are generated and structured in a similar way.

6.3.2 Pitch adaptation

As described in the conversion rules (chapter 5), the pitch is adapted during syllables of words in the fragment that are sentence accents. We will not explain the formula that is used for pitch manipulation again here, but only describe the way the formula is used in *Praat*. The formula is the following:

$$y'(t) = \begin{cases} y(t).(1+(\sin((((t-t_1)/(t_2-t_1))m_2 pi)+m_1 pi)/n)) & ,if \quad t \in [t_1,t_2] \\ y(t) & ,else \end{cases}$$

With the following variables and constants:

y	acoustic feature value y as function of t
y'	manipulated pitch values
m_1	constant determining starting point of sine waveform that should be used
m_2	constant determining fraction of sine waveform that should be used
n	constant determining the degree of adaptation

Each single pitch adaptation is carried out on a specific time domain, indicating a certain syllable of an accented word. Using the pitch script all pitch adaptations for a certain fragment can be carried out subsequently. First, the pitch tier of a certain fragment must be extracted from the original fragment. During the manipulation this tier is adapted and afterwards the original tier is replaced by the manipulated tier.

In order to perform a pitch adaptation on the pitch tier for each syllable the following *Praat* command is executed:

```
Formula...
 if x > syll_start then
  if x < syll_end then
   self * (1+( sin(((x-syll_start)/syll_leng)*0.75*pi) / (avg_pitch / dmpi )))
  else
   self
  fi
 else
  self
 fi
```

The parts of the script that are in italics are the parts where values for each specific syllable must be provided:

syll_start	start of syllable in seconds
syll_end	end of syllable in seconds
syll_length	duration of syllable in seconds (`syll_end-syll_start`)
avg_pitch	average pitch of the speech fragment that is manipulated
dmpi	desired maximum pitch increase of the manipulation

Before the script is run for each syllable those syllable specific values are determined. So in the case we want to manipulate a certain syllable with time domain [`0,71 , 0,92`] and we want to have a maximum pitch increase of 40 Hz, knowing that the average pitch of the fragment is 105 Hz, we must execute the following script:

```
Formula… if x > 0,71 then if x < 0,92 then
  self * (1+( sin(((x-0,71)/0,21)*0.75*pi) / (105 / 40 )))
else self fi else self fi
```

The final script for a certain sentence consists of a pitch formula for each of the sentence accent syllables and can be executed at once for the entire sentence.

6.4 Evaluation environment

6.4.1 Introduction

We will to try to computerise the evaluation process as much as possible. Of course this approach should not result in misinterpretation by the participant of the goals of the evaluation, for this purpose the introduction to the experiment will always be conducted by the researcher personally. The automation of the process is primarily aimed at the automatic storage of answer data, such that this data can be analysed more rapidly.

The evaluation environment must support the following functionalities:

- Textual introduction (Appendix F) to participant (supported by researcher in person), after this introduction the user must be able to perform the evaluation independently.
- Supply the participants with a series of fragments that one by one can be listened to several times, supply forced response questions with possible answers.
- Automatic storage of the answers given by participant, including the number of times the participant listened to a certain fragment.
- Supply several well-organised data views in which the researcher can easily look at the results of the evaluation. The views can be used as an input for statistical software.
- Administrative researcher functions like simply add, remove, replace fragments and change order of fragments (since we have to do three evaluations with multiple test sets).

In order to fulfil these requirements a web-application is developed. The application core is a database in which the questions, fragments and answers are stored. The following paragraph will briefly describe the application.

6.4.2 Application

The application uses a MySQL database to store data; the application itself is implemented in an HTML user interface. PHP scripting is used to realise the linking of the user interface with the database. An advantage of the web based approach is that the evaluation can take place from any location in which a PC with internet connection is available.

Three kinds of information are important in the evaluation process: participant, question and answer data. The database of the application has a table for each these three data groups. The schema underneath shows the structure of the database and the fields that each of the tables contains (figure 6.1).

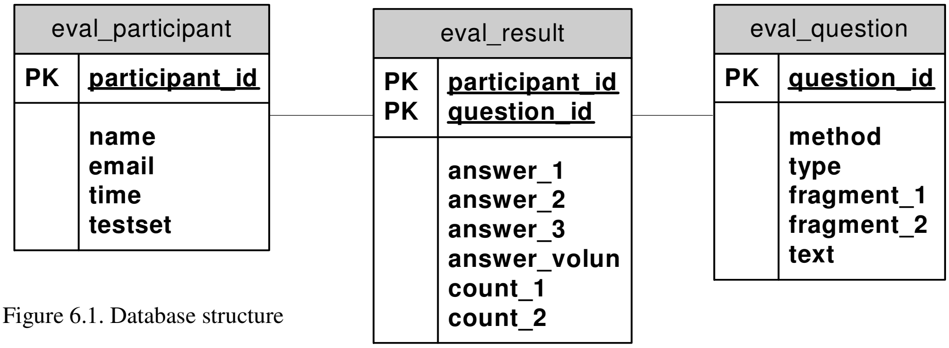


Figure 6.1. Database structure

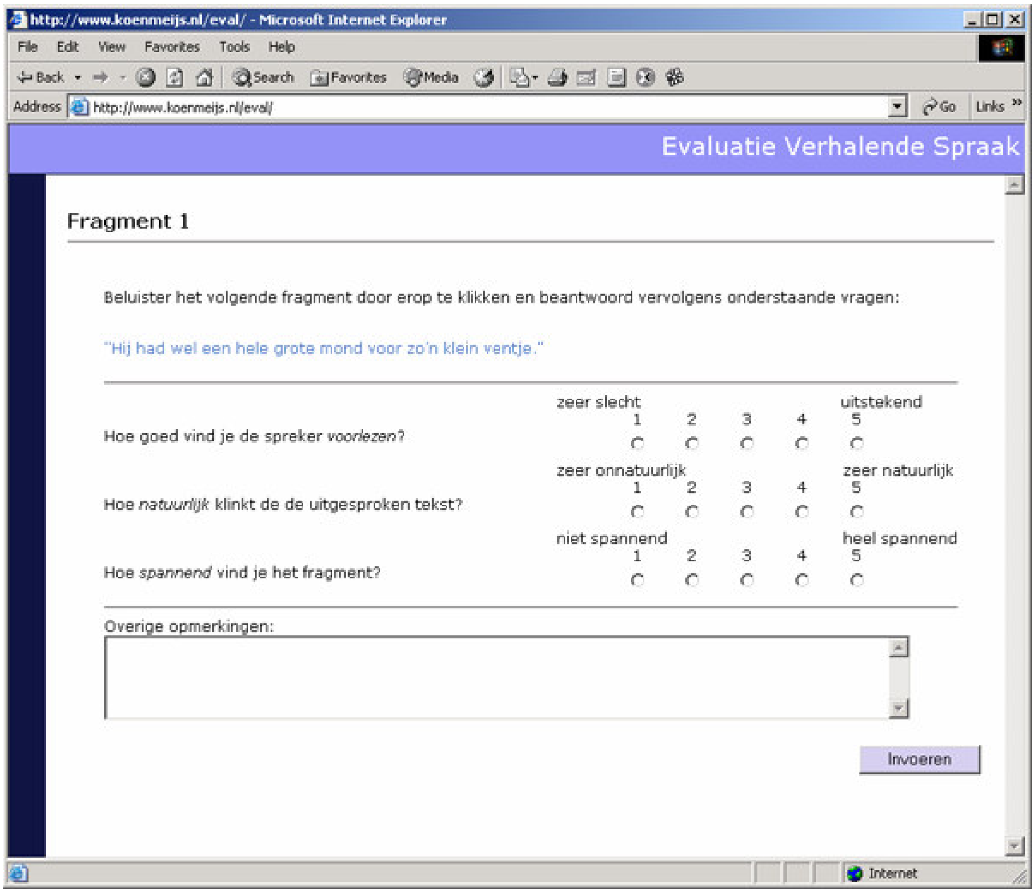
The procedure of the application is as follows. Before the evaluation is carried out the researcher adds questions to table *eval_question*. Is this way the application can determine which questions should be asked and how each question should be supplied to the participant. When the evaluation starts the participant has to provide some personal information which is stored in *eval_participant*. Next the application will one by one fetch the questions from table *eval_question.* The answers that are given for each question will be stored in *eval_result*. After all questions have been processed the evaluation is finished.

We will now discuss the tables in more detail. Table *eval_participant* contains information about the participants. Besides normal personal information like name and e-mail address this table contains the number of the test set that was supplied to the participant (in the case of the conversion rule evaluation there's only one test set). Each participant that is inserted into the database is assigned a unique participant identification number (primary key *participant_id*), this id is also used in table *eval_result* to separate the results of different participants.

Table *eval_question* specifies the properties of questions that are shown to the participant. First, each question is stored uniquely by *question_id* (primary key). Each question involves one or two sound fragments, which are specified in fields *fragment_1* and *fragment_2.* The values of the fields are in fact filenames of audio files that are on the file system. The number of fragments that are supplied to the participant depend on the method of evaluation that is used for a certain question (binary choice or Likert method, see 6.2.2) and is specified in field *method.* Each question concerns one of the two possible natures of fragments (narrative style or climax), which is specified in field *type.* In order to show the participant the text of the spoken fragments this must be stored as well, which is done in field *text.*

The *eval_result* table contains answer values for each unique combination of *question_id* and *subject_id* (together forming primary key). Depending on the question method that is used, there are one or three answers to store for each question. The answers are stored in fields *answer1, answer2* and *answer3*. Each question that is supplied to the participant is accompanied by an input field in which the user can type general remarks about the fragments; this voluntary remark data is stored in field *answer_volun*. The table also contains two fields (*count_1* and *count_2*) in which for the fragments the number of times they were listened to is tracked.

Figure 6.2 shows a screenshot of the application in progress.


http://www.koenmeijs.nl/eval/ - Microsoft Internet Explorer
File Edit View Favorites Tools Help
Back Search Favorites Media
Address http://www.koenmeijs.nl/eval/
Go Links
Internet

Evaluatie Verhalende Spraak

Fragment 1

Beluister het volgende fragment door erop te klikken en beantwoord vervolgens onderstaande vragen:

"Hij had wel een hele grote mond voor zo'n klein ventje."

	zeer slecht 1	2	3	4	uitstekend 5
Hoe goed vind je de spreker *voorlezen*?	○	○	○	○	○
	zeer onnatuurlijk 1	2	3	4	zeer natuurlijk 5
Hoe *natuurlijk* klinkt de de uitgesproken tekst?	○	○	○	○	○
	niet spannend 1	2	3	4	heel spannend 5
Hoe *spannend* vind je het fragment?	○	○	○	○	○

Overige opmerkingen:

Invoeren

Figure 6.2. Screenshot of evaluation application

7 Constant evaluation

7.1 Introduction

The goal of the constant evaluation is to determine the best values for the constants used in the conversion rules (chapter 5). The procedure used in the constant evaluation is described in the evaluation plan (§6.2); here we will describe the stimuli and questions that were asked (§7.2), discuss the constant value ranges on which the stimuli are based (§7.3) and give the results of the evaluation (§7.4). The last paragraph of the chapter contains the conclusions of the constant evaluation.

7.2 Stimuli and questions

The constant evaluation is intended to determine the best pitch, intensity and duration constant values of the conversion rules. For each constant we will create pairs of fragments based on different values of the constant, while keeping the other constants at the same value. Besides providing pairs of two fragments with differing constant values we will also include neutral fragments, which are used as a comparison baseline.

The fragments will be provided to one group of participants. A participant has to choose the best fragment based on a certain criterion. The number of fragments that are created depends on the number of free constants we have. The following fragments form the types of fragments, with the number of fragments indicated between parentheses:

- Sentence spoken in narrative style (12)
- Sentence spoken in narrative style containing sudden climax (5)
- Three successive sentences spoken in narrative style containing increasing climax (6)

Regarding the order of the fragments, the general procedure is that the constants are handled one by one and used as a basis for a fragment. For every new question the first fragment is used as a baseline to compare the second to. The second then has a new value for one of the evaluation constants.

We need to evaluate the sound fragments based on certain judgement criteria. The most important criterion to evaluate is the naturalness of the speech. Another criterion is the tension the participant perceives in a fragment. Depending on the nature of the fragment (narrative style or climax) we will ask one of the following questions (table 7.1, also showing the question in Dutch as it is asked in the evaluation application):

Question	Answers
“Welke van deze fragmenten vind je het meest natuurlijk klinken?” (“Which of the fragments do you perceive as the most natural sounding?”)	• ‘fragment 1’ • ‘geen verschil’ (‘no difference’) • ‘fragment 2’

“In welke van deze fragmenten vind je de spanning het beste weergegeven?” (“Which of the fragments has the best expression of tension?”)	• ‘fragment 1’ • ‘geen verschil’ (‘no difference’) • ‘fragment 2’

Table 7.1. Constant evaluation questions

7.3 Creation of fragments

We will not one by one discuss all fragments which are used in the constant evaluation, but summarise the variations in constant values we used to create the fragments. A fully detailed description of the fragments can be found in appendix B.
The best value of some of the constants in the conversion rules have not been determined yet, the constants for which this is the case can be seen in table 7.2 below. Some adaptations were made to the lower boundaries of the constant values, therefore there are two range columns in the table. The reason of this adaptation is explained below.

<table>
<tr><th>Phenomenon</th><th>Property</th><th>Constant</th><th>Description</th><th>Range (original)</th><th>Range (adapted)</th></tr>
<tr><td>narrative style</td><td>pitch</td><td>m_1</td><td>Starting point of fraction of the sine curve to use</td><td>0 or 0,25</td><td></td></tr>
<tr><td>narrative style</td><td>pitch</td><td>m_2</td><td>fraction of the sine curve to use</td><td>0,75 or 0,5</td><td></td></tr>
<tr><td>narrative style</td><td>pitch</td><td>desired_max_pitch_increase</td><td>Maximum increase with respect to the average pitch</td><td>40 - 90 Hz</td><td>30 - 90 Hz</td></tr>
<tr><td>narrative style</td><td>intensity</td><td>c</td><td>Intensity value increase</td><td>4 - 7 dB</td><td>2 - 6 dB</td></tr>
<tr><td>narrative style</td><td>intensity</td><td>k</td><td>time domain increase</td><td>0 - 0,2 sec</td><td></td></tr>
<tr><td>narrative style</td><td>tempo</td><td>-</td><td>global tempo</td><td>3,0 – 3,6 s.p.s.</td><td></td></tr>
<tr><td>narrative style</td><td>vowel duration</td><td>-</td><td>Accented vowel duration increase factor</td><td>1 or 1.5</td><td></td></tr>
<tr><th>Phenomenon</th><th>Property</th><th colspan="2">Description</th><th>Range</th><th>Range</th></tr>
<tr><td>climax 1*</td><td>pitch</td><td colspan="2">increase of pitch at start of climax</td><td>80 - 120 Hz</td><td></td></tr>
<tr><td>climax 1</td><td>pitch</td><td colspan="2">behaviour of pitch after pitch rise</td><td>stay constant or increase 200Hz</td><td></td></tr>
<tr><td>climax 1</td><td>intensity</td><td colspan="2">increase of intensity at start of climax</td><td>6 - 10 dB</td><td></td></tr>
<tr><td>climax 1</td><td>intensity</td><td colspan="2">behaviour of intensity after intensity rise</td><td>stay constant or decrease</td><td></td></tr>
<tr><td>climax 2*</td><td>pitch</td><td colspan="2">behaviour of pitch contour during climax</td><td>start at + 100Hz
top at + 130Hz</td><td>start at + 25-50Hz
top at + 60-80Hz</td></tr>
<tr><td>climax 2</td><td>duration</td><td colspan="2">maximum accented vowel duration increase factor during climax</td><td>1,5 - 2</td><td></td></tr>
</table>

* climax 1 = sudden climax, climax 2 = increasing climax

Table 7.2. Constant value ranges

In order to determine the best values we provided the participant several variations of fragments. One approach was to let the participant compare an original Fluency fragment with a manipulated one. The other approach was to let the participant judge two manipulated fragments differing in the value of one of the constants, while keeping the other constants at the same value in both fragments. By first comparing a neutral fragment to a manipulated fragment based on a certain constant value, and then comparing the same manipulated fragment to another manipulated fragment with a different value for the constant in question, we can see how the constant difference is in the proportion of the neutral fragment.

During the production of the fragments containing narrative style, it turned out that some of the constant ranges that were found in the analysis were not appropriate for synthetic speech. After we applied the pitch and intensity values that were observed for storytellers to the *Fluency* speech fragments, it turned out that those fragments sounded too unnatural to even consider including in the evaluation. This phenomenon occurs most when using relatively high pitch and intensity values. Only the constant values that are near the lower bounds of the range sounded acceptable. For this reason we decided to shift down the ranges of some constants for which this applies, which is already indicated in table 7.2. For narrative style, the lower pitch bound is now 30Hz instead of 40Hz, the lower intensity bound is now 2dB instead of 4dB. For the increasing climax the pitch start increase is 25-50Hz, the maximal increase is 60-80Hz.

One possible explanation for this phenomenon is that the post-processing (processing of fragments after *Fluency* has created them) using *Praat* is based on certain methods in which information is lost. In order to perform a pitch adaptation on a certain fragment, a pitch contour analysis is conducted based on a fixed size time window and a certain frequency domain. After this the pitch contour can be adapted (which is nothing more than a simple arithmetic operation) followed by a resynthesis operation using PSOLA (§2.6.2) which results in a manipulated sound object. A lot of processing of the sound is done here, sometimes resulting in a decreased quality of speech. It turns out that the quality decrease only takes place when an actual adaptation with relatively high values is performed. If we only perform the pitch contour derivation afterwards followed by the resynthesis without adapting any pitch values, it turns out that there is no audible quality loss. So the cause of the quality loss must be in the combination of adapting the pitch with high values and afterwards applying the resynthesis, which is a known problem of synthesis algorithms (§2.6.2). This is confirmed in [12] where is stated that "signal processing inevitably incurs distortion, and the quality of speech gets worse when the signal processing has to stretch the pitch and duration by large amounts".

One could argue that for a reliable evaluation it is necessary to have all fragments used in the evaluation processed by *Praat*, so both fragments without narrative style and fragments containing narrative style should be processed. In this way all fragments get the same treatment and if any quality loss is present, this is present in all fragments. The problem here is that the cause of the quality loss is not in the pitch contour derivation and resynthesis only, but also in the combination of those with a high pitch adaptation. So to treat all fragments equally, the neutral fragments should also be pitch manipulated, which is a contradiction. One could then argue that the pitch of a neutral fragment should be adapted by a certain amount, and then perform the exact inverse pitch adaptation so the original pitch value is obtained. First there is a problem here that it isn't possible to determine how large this adaptation should be; moreover this adaptation doesn't introduce the desired result because we are only performing arithmetic operations on absolute pitch values inside a certain time domain. No synthesis is performed during this process, so we can change the pitch values to whatever value, as long as we return them to their original, there will be no influence on the result.

Summarising, quality loss only takes place in the case of a high pitch value increase combined with resynthesis. Processing neutral fragments with *Praat* is not necessary, because the quality of the neutral fragments is not affected by the operations that would be performed.

7.4 Results and discussion of the constant evaluation

The average evaluation answer values given by the participants are included in appendix C. In table 7.3 underneath are the best constant values that were derived from the results of the evaluation.

<table>
<tr><th>Phenomenon</th><th>Property</th><th>Constant</th><th>Description</th><th>Range</th><th>Result</th></tr>
<tr><td>narrative style</td><td>pitch</td><td>m_1</td><td>Starting point of fraction of the sine curve to use</td><td>0 or 0,25</td><td>0,25</td></tr>
<tr><td>narrative style</td><td>pitch</td><td>m_2</td><td>fraction of the sine curve to use</td><td>0,75 or 0,5</td><td>0,5</td></tr>
<tr><td>narrative style</td><td>pitch</td><td>desired_max_pitch_increase</td><td>maximum increase with respect to the average pitch</td><td>30 - 90 Hz</td><td>40 Hz</td></tr>
<tr><td>narrative style</td><td>intensity</td><td>c</td><td>intensity value increase</td><td>2 - 6 dB</td><td>2 dB</td></tr>
<tr><td>narrative style</td><td>intensity</td><td>k</td><td>time domain increase</td><td>0 - 0,2 sec</td><td>0 sec</td></tr>
<tr><td>narrative style</td><td>tempo</td><td>-</td><td>global tempo</td><td>3,0 – 3,6 s.p.s.</td><td>3,6 s.p.s.</td></tr>
<tr><td>narrative style</td><td>vowel duration</td><td>-</td><td>accented vowel duration increase factor</td><td>1 or 1.5</td><td>1.5</td></tr>
<tr><th>Phenomenon</th><th>Property</th><th colspan="2">Description</th><th>Range</th><th>Result</th></tr>
<tr><td>climax 1</td><td>pitch</td><td colspan="2">increase of pitch at start of climax</td><td>80 - 120 Hz</td><td>80 Hz</td></tr>
<tr><td>climax 1</td><td>pitch</td><td colspan="2">behaviour of pitch after pitch rise</td><td>stay constant or increase 200Hz</td><td>stay constant</td></tr>
<tr><td>climax 1</td><td>intensity</td><td colspan="2">increase of intensity at start of climax</td><td>6 - 10 dB</td><td>6 dB</td></tr>
<tr><td>climax 1</td><td>intensity</td><td colspan="2">behaviour of intensity after intensity rise</td><td>stay constant or decrease</td><td>decrease</td></tr>
<tr><td>climax 2</td><td>pitch</td><td colspan="2">behaviour of pitch contour during climax</td><td>start at + 25-50Hz
top at + 60-80Hz</td><td>start at +25Hz, gradual rise to max 60 (fragment 17)</td></tr>
<tr><td>climax 2</td><td>duration</td><td colspan="2">maximum accented vowel duration increase factor during climax</td><td>1,5 - 2</td><td>1,5</td></tr>
</table>

Table 7.3. Best constant values

One of the difficulties in judging the evaluation results is that each participant has his own personal preferences regarding naturalness and tension, which in some cases causes participants to have contradictory opinions about fragments. In spite of the influence of subjectivity, we can still draw conclusions concerning the best values for the constants, because there is also agreement in a lot of cases. The most salient evaluation results will be discussed here.

It was remarkable that during the evaluation of the pitch characteristics of narrative style almost nobody noticed the variations in constants m_1 and m_2. The few cases that the listener noticed a difference he couldn't describe it, and didn't prefer any of the fragments. The values we choose will be discussed below.

With regard to intensity constants, in general the participants only notice that the fragments sound louder, but they don't notice the differences in position and length of the intensity increase. So although the intensity is only increased surrounding the accented syllable participants don't notice the locality of it.

In the case that participants don't prefer one fragment over the other (mean answer value near expectation value with low standard deviation) or in the case that the evaluation results were highly diverging among participants (high standard deviation) we have made choices. One example of the first case is the judgement of the form of the pitch contour (determined by m_1 and m_2). Because participants don't notice any difference between the two pitch contour variants we choose the second variant, based on the fact that in the analysis it turned out that this variant occurs most.

The preferred intensity value for narrative style is an example of divergence among the answers. Some answers show that participants don't like large intensity increases (4dB or 6dB), but there is discord about the best value (no intensity increase or 2 dB). The discord is also visible in the number of times the fragments of the question regarding this intensity increase (0 or 2 dB, question 8, appendix C) were listened to. On average fragment 1 was listened to 2,2 times, fragment 2 was listened to 1,8 times, while the average of times that all questions are listened to is 1,5. So these fragments were listened to more than average, which is supposedly caused by the difficulty of judgement. Additional remarks in the evaluation show that participants don't really believe that the intensity increase contributes much to the naturalness of the fragment. But neither do they experience it as unnatural. Based on this we have once again returned to the analysis and based our choice on the fact that in general there is a intensity increase observable, so we choose the intensity value of 2 dB.

The determination of the best values for the climax constants is quite straightforward because there were more unanimous preferences for certain values. The best values turned out to be those near the lower bound of the constant ranges, higher values were experienced unnatural or 'too much'. One choice has to be made because the evaluation didn't give a solution. In the case of increased duration during climax 2 participants didn't really notice any difference (question 18) or answers were diverse (question 20). Because the increase isn't rejected we will stick to the analysis results, meaning we will include the duration change but use an average value (factor 1,5).

7.5 Conclusion

In general can be concluded that the participants experience the manipulations as an improvement, but some of the fragments used in this evaluation sounded unnatural because the amount of manipulation was exaggerated. Based on this we chose the constant values in such a way that they are appreciated maximally.

From the free responses that were supplied by participants it also turns out that participants consider the application of the rules to be contributing to the quality of storytelling, although the quality of storytelling was not a judgement criterion in this evaluation.

8 Conversion rule evaluation

8.1 Introduction

After the best values for the constants in the conversion rules have been determined, we have a deterministic set of conversion rules, meaning that a certain unique input will always give the same output. The rules are now ready to be applied to neutral speech and should yield narrative speech. The conversion rule evaluation is intended to determine the added value of the application of the conversion rules, which means we want to find out whether the application really contributes to the quality of storytelling.

Since the results of the constant evaluation already indicated that participants consider the rules to be contributing to the narrative quality, this evaluation has an affirmative role. The starting point in this evaluation is that we don't need statistically significant proof for the increase of narrative quality, so we can keep the evaluation relatively small (the implementation evaluation will be more extensive). So only a small set of fragments is evaluated and the number of participants is limited (eight persons).

In this chapter we will describe the questions and stimuli that are used in the evaluation (§8.2), formulate a hypothesis (§8.3) and describe and discuss the results of this evaluation (§8.4). The chapter ends with a conclusion (§8.5).

8.2 Questions and stimuli

The series of fragments will be evaluated based on the 5-scale method by Likert [11]. We want to evaluate the speech fragments based on certain judgement criteria. The criterions to base the evaluation on are the quality of the storytelling, the naturalness of the speech and the amount of tension the participant experiences in a fragment. Based on this we will accompany each fragment with three questions and ancillary answers (table 8.1):

Question	Answer range
"Hoe goed vind je de spreker voorlezen?" ("How do you judge the quality of storytelling of this speaker?")	• 1 = 'zeer slecht ('very bad') • … • 5 = 'uitstekend' ('excellent')
"Hoe natuurlijk klinkt de uitgesproken tekst?" ("How do you judge the naturalness of the speech?")	• 1 = 'zeer onnatuurlijk' ('very unnatural') • … • 5 = 'zeer natuurlijk' ('very natural')
"Hoe spannend vind je het fragment?" ("How tense do you perceive the fragment?")	• 1 = 'niet spannend' ('not tense') • … • 5 = 'heel spannend' ('very tense')

Table 8.1. Conversion rule evaluation questions

The first question is quite straightforward and is intended to determine the storytelling quality of the fragment in question. Since for each sentence that is spoken both a neutral and a manipulated fragment are included in the test sets we can compare the differences in judgements of the two versions.

The second question is included because we wanted to separate the quality of storytelling from the naturalness of the sentence. It is conceivable that although a participant experiences a fragment of being of good storytelling quality, he doesn't consider it very natural. By separating the two aspects we are able to obtain more elaborate information. The first two questions can be a source of confusion, because participants might have problems separating the two. To avoid confusion the two questions will be explained to the participants before the start of the evaluation.

The last question that is asked is aimed at measuring the effect of the addition of climaxes. Because a lot of fragments are not intended to have an explicit tension course and those containing a climax do so, we can compare the two groups and see how much the climax contributes to the tension course of the speech.

All stimuli that are created have fixed constant values as determined in the constant evaluation (§7.4, table 7.3) and the text that is used in the stimuli is taken from child stories, but any fairy tale referencing elements are removed or replaced.

The total set of stimuli consists of *sixteen* stimuli, of which eight are neutrally spoken unique text fragments and eight are the same unique text fragments spoken in narrative style or climax. These last eight stimuli can be divided in five containing narrative style only, and three containing both narrative style and climaxes. We will create two test sets across which we divide the sixteen stimuli; the first test set is to be evaluated by one half of the participant group, the other set by the other half. We will divide the stimuli such, that each unique text fragment is presented in neutral form to one group, and the same fragment in manipulated form to the other group. The following schema summarises the above (figure 8.1):

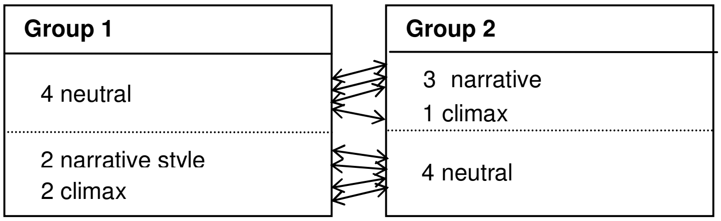


Figure 8.1 fragment division

For the evaluation of increasing climaxes we will use stimuli that are three sentences long. A problem with shorter stimuli is that they may be too short for the participant to judge, because the desired amount of tension can not be communicated to the participant in a short stimulus.

The sixteen stimuli will be divided over two groups of participants of each 4 participants. The eight stimuli of each group will be provided in random order such that there is no possible guessing of nature of the fragment (narrative or climax).

The list of stimuli is provided in full detail in appendix D.

8.3 Hypothesis

The null hypothesis is that for a certain fragment, the manipulated version is not rated significantly better than the neutral version with respect to narrative quality, naturalness and tension display. So the average rating of the three judgement aspects of both versions is expected to be equal. We will reject this hypothesis if there is valid proof that the two are not equal.

8.4 Results

8.4.1 Introduction

Before we started the evaluation we were aware of the fact that some phenomena could influence the judgement by the participants. First there is the problem that although before the experiment we emphasise that people shouldn't judge the intelligibility or sound quality of artificial speech but look at the narrative aspect of it, it turns out to be very hard for participants to separate those aspects.

Another phenomenon that appeared during the evaluation is that participants don't agree with the position of the accents in the sentence. All sentences used in the evaluation were taken from original fairy tales. The first group of fragments (a-fragments) is only pronounced by *Fluency,* the second group (b-fragments) is manipulated in sentence accent positions in order to create narrative style. So all accents in the b-fragment group correspond to those realised by storytellers. Some participants though experience these positionings as completely unnatural and as a consequence judge the fragment analogously. So there turns out to be a large amount of subjectivity in the judgement of the accent positioning, which can't be avoided because of its unpredictability.

A third phenomenon that turned out during the evaluation was that although it was the setup of the evaluation to include neutral fragments and have them judged too, participants sometimes tended to be looking for narrative properties in fragments in which they were not present. This phenomenon was observed in some of the free responses that participants gave. As a consequence they overrated the neutral fragments, believing 'there should be something'.

Because the constant evaluation already yielded that participants prefer the narrative versions of the fragments, we decided to keep the evaluation small. So only a small set of fragments was evaluated and the number of participants was limited (§8.1). This small setup and the uncertainty introduced by the observed phenomena contribute to the decision to describe the analysis of the evaluation results in a qualitative way. Analysing them purely statistically is undesirable, because the statistical analysis methods require a certain degree of value variation and dataset size to draw conclusions with an acceptable significance. So we will base our conclusions on observations like the mean answer value, the standard deviation, range, etc. Still we will apply two methods that determine whether there exists any statistical difference in average between the two fragments. Those two methods will be briefly explained in the next paragraph.

8.4.2 Statistical methods

In the statistical processing of the results it is important to realise that it is not enough only to compare the average answer values of a neutral and manipulated fragment. If we want to see whether there is significant difference in the judgement of the fragments, we have to take into account the variability around the means. A small difference between means is easily detectable if variability is low. In order to test the variability we will use the t-test, a well known method for this purpose [11]. The t-test calculates the t-value, a value which is a factor representing the ratio between the difference in mean of two datasets and the standard error of the difference between the sets. After this value has been calculated the value can be looked up in a table of significance to test whether the ratio is large enough to say that the difference between the sets is not likely to have been a chance finding.

The other method we will use is the Mann-Whitney U test. This test is based on combining and ranking all values of the datasets and adding all ranks for both sets. The number of times a score from group 1 precedes a score from group 2 and the number of times a score from group 2 precedes a score from group 1 are calculated. The Mann-Whitney U statistic is the smaller of these two numbers. This value can be looked up in a significance table resulting in a conclusion whether the sets are equal or differing with certain significance. Compared to the t-test the Mann-Whitney U test is stricter so it is less likely to produce significant results in our case.

8.4.3 Statistical results and discussion

In this section for each question that was asked in the evaluation the results are given and interpreted. The following tables show the mean, standard deviation and value range of the answer values (on a 1 to 5 scale) of the first question ("How do you judge the quality of storytelling of this speaker?"). Table 8.2a shows these statistics for non-manipulated fragments (indicated by "**a**"), table 8.2b for manipulated fragments (indicated with "**b**"). Fragments *1* until *5* only contain narrative style; *6* until *8* contain both narrative style and climax. We also applied the t-test and Mann-Whitney method to each couple of *a/b*-fragments. The significance with which can be said that the two series of results are statistically different is listed in table 8.2c.

Fragment	**1a**	**2a**	**3a**	**4a**	**5a**	**6a**	**7a**	**8a**
mean	2,5	2,75	3,25	2,75	2	1,75	3,25	2,5
standard deviation	1,91	0,96	0,96	0,50	0,82	0,96	0,50	0,58
range	4	2	2	1	2	2	1	1

Fragment	**1b**	**2b**	**3b**	**4b**	**5b**	**6b**	**7b**	**8b**
mean	2,25	2,75	3,75	3,25	3,5	3,75	3,5	4
standard deviation	1,50	0,96	0,50	0,96	1,00	0,96	0,58	0,82
range	3	2	1	2	2	2	1	2

Fragment	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**
t-test significance	0,84	1,00	0,39	0,39	0,06	0,03	0,54	0,02
mann-whitney. significance	0,88	1,00	0,41	0,35	0,07	0,04	0,50	0,04

Table 8.2 a,b,c. Statistics for question "How do you judge the quality of storytelling of this speaker?"

It is clearly visible that the first four narrative fragments are not considered significantly better than the original fragments. The fifth fragment however has more divergent mean values for the two versions, though still not with statistically acceptable significance (significance lower or equal to 0,05 is generally considered acceptable). Solely based on mean value there may be concluded that participants regard the manipulations as contributions to the quality of storytelling, because in general the mean value is equal or higher for the manipulated fragment. Looking at the voluntary remarks people gave during the first four manipulated fragments it is striking that in all four cases people criticise the positions of the sentence accents and therefore rate the fragments lower. So although the accent positions were taken from real storyteller fragments, participants don’t always accept these positions.

Taking a look at the quality of storytelling of fragments containing a climax, fragment *6b* and *8b* are significantly tenser than their neutral equivalent. This is clearly visible in the mean values of the fragments and the mean values are proven to be statistically different by both tests. So although the dataset is small it is possible to get significantly different results.

Fragment *7a* and *7b* have about equal mean value, but for *7a* the mean value itself is relatively high compared to that of fragment *6* and *8*. So it seems that the original Fluency pronunciation already has a higher degree of storytelling quality.

Tables 8.3 a,b,c show identical statistics for the question “How do you judge the naturalness of the fragment”:

Fragment	**1a**	**2a**	**3a**	**4a**	**5a**	**6a**	**7a**	**8a**
mean	2,5	3	3	2,5	1,75	1,75	3,25	2,75
standard deviation	1,00	0,82	0,82	1,00	0,96	0,96	0,50	0,96
range	2	2	2	2	2	2	1	2

Fragment	**1b**	**2b**	**3b**	**4b**	**5b**	**6b**	**7b**	**8b**
mean	2	2,5	2,75	2,5	3,5	3,25	3	4
standard deviation	1,41	0,58	0,50	1,00	1,00	0,96	0,82	0,00
range	3	1	1	2	2	2	2	0

Fragment	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**
t-test significance	0,59	0,36	0,62	1,00	0,05	0,07	0,62	0,04
Mann-Whitney. significance	0,35	0,34	0,62	0,76	0,05	0,08	0,62	0,05

Table 8.3 a,b,c. Statistics for question “How do you judge the naturalness of the fragment?”

Looking at the mean values only, once again the first four questions show that manipulated fragments are considered less natural. The general opinion about these fragments by the participants was that they sounded unnatural because of the accent positioning. As a result, especially in the case of this question where the naturalness is judged, this opinion is reflected in the answers.

Fragment *5b* is regarded better than *5a*, with a mean difference of 1,75 points. Contrary to the answers for this fragment regarding quality of storytelling, this time the difference can be statistically proven with an acceptable significance of 0,05 for both test methods.

Considering the results for climaxes, the mean value for fragment *6b* is substantially higher than that of fragment *6a*, almost proven with acceptable significance (0,07). The same goes for fragment *8*, but this time for both test methods there is prove with significance. Fragment *7* scores high again in both versions, leading to the similar conclusion as for the previous question, namely that the Fluency pronunciation is already quite natural.

One important general observation is that there is a lot of agreement between the answers for the first and second evaluation question. So there seems to be a relation between narrative quality and naturalness. From the results is turns out that fragments of low storytelling quality are judged correspondingly on naturalness. The same goes for fragments of high storytelling quality, they are also judged high on naturalness. The relation is one-directional, because a very natural sounding fragment doesn't have to be of good storytelling quality.

The last series of tables (table 8.4 a,b,c) shows the statistics for the question "How tense do you experience the fragment?":

Fragment	**1a**	**2a**	**3a**	**4a**	**5a**	**6a**	**7a**	**8a**
mean	1,25	1,75	2,25	2	1,5	1,25	2,25	2
standard deviation	0,50	0,96	1,50	0,82	0,58	0,50	0,50	1,15
range	1	2	3	2	1	1	1	2

Fragment	**1b**	**2b**	**3b**	**4b**	**5b**	**6b**	**7b**	**8b**
mean	2,75	2,5	1,75	2,25	1,75	3,25	3,5	4
standard deviation	1,26	0,58	0,50	0,96	0,96	0,50	0,58	0,82
range	3	1	1	2	2	1	1	2

Fragment	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**
t-test significance	0,07	0,23	0,55	0,71	0,67	0,001	0,02	0,03
Mann-Whitney. significance	0,09	0,22	0,76	0,65	0,75	0,02	0,03	0,04

Table 8.4 a,b,c. Statistics for question "How tense do you experience the fragment?"

The goal of this question is to find out whether the presence of climaxes influences the tension experience of participants. The first five fragments in which no climax was included, for both the *a* and *b* versions no significant difference is observed, although most narrative style adapted fragments are considered more tense. Taking a look at the climax fragments, for all fragments there is a significant difference in the judgement of tension, the significance being relatively high in all cases. So from this we may conclude that participants believe the climaxes to be contributing positively to the amount of tension that is experienced.

One last statistical analysis that is conducted is to see the difference in judgement between the two participant groups. Because the size of the group is relatively small, it is possible that the two participant groups have diverging average judgements purely based on the constitution of the groups. If this kind of bias would be present this could be a problem because then it's hard to say

whether this bias is caused by the group constitution or whether this difference is solely stimuli related (so not being a bias at all). On the other hand, if we can prove that the average judgements of the neutral and manipulated fragments for the both groups are not significantly different, then we know with more certainty that our results are unbiased, which is of course desirable.
The following table (table 8.5) shows the mean judgements of each participant group separated by fragment group and question.

participant group	1						2					
fragment group	a-fragments			b-fragments			a-fragments			b-fragments		
question	1	2	3	1	2	3	1	2	3	1	2	3
mean	2,33	2,41	1,41	3,60	3,15	2,65	2,75	2,65	2,00	2,91	2,58	2,83

Table 8.5. Group separated mean judgements

There are some mean differences observable between the groups, so the next step is to calculate with the t-test whether the judgement differences between groups are significant. The t-test is the less strict of the two difference tests, so of this method doesn't find any significant difference, the other wouldn't either, meaning the group judgements are not significantly different.
So we compared the differences in means for all neutral fragments of group 1 with all neutral fragments of group 2, we did the same for the manipulated fragments (table 8.6).

fragment group	a-fragments			b-fragments		
question	1	2	3	1	2	3
significance	0,33	0,51	0,07	0,06	0,12	0,64

Table 8.6. Group separated difference significance levels.

For both the neutral and the manipulated fragments, for neither of the three questions there is a significant difference between the average judgements of the groups (all significance above 0,05). But for two questions there is a near significant value (0,07 and 0,06), which means there's almost a significant difference between the average answer values of the two groups. This is the case for the judgement of tension of the a-fragments and the judgement of storytelling of the b-fragment group.

8.5 Conclusion

Looking in a qualitative way at the results only, regarding quality of storytelling we can say that manipulated fragments are judged equal or better, especially in the case of climaxes. For two of the three climax fragments this can be proven with statistically acceptable significance. Based on this we may conclude that the addition of narrative style in some cases increases the appreciation of the storytelling, especially in the case of climaxes.
The judgement of naturalness shows a corresponding pattern, but in general participants don't judge the manipulated fragments more natural (sometimes even the opposite) than the original ones. Once again the climax fragments are appreciated higher.
The appearance of climaxes in fragments is really contributing to the amount of tension that is experienced. All climax fragments are considered to be significantly tenser.

The controversial positioning of the sentence accents turns out to be an important factor of negative influence on the judgement of fragments, which is mainly perceptible in the judgement of naturalness and storytelling quality.

Regarding group bias, we proved that there are no significant differences between the judgements of the two groups, but in two of the six combinations of fragment groups and questions there is almost a significant difference. It is hard to say what's the cause of this, this can be a group bias or it can be contributed to the nature of the fragment itself.

9 Implementation

9.1 Introduction

In this section we will describe the module that has been implemented in order to create narrative speech automatically. In the project description and goals section of this report's introduction (§1.1) we have already briefly sketched the steps that have to be taken in the implementation to automatically generate narrative speech. We will repeat the schematic representation of the process here in figure 9.1.

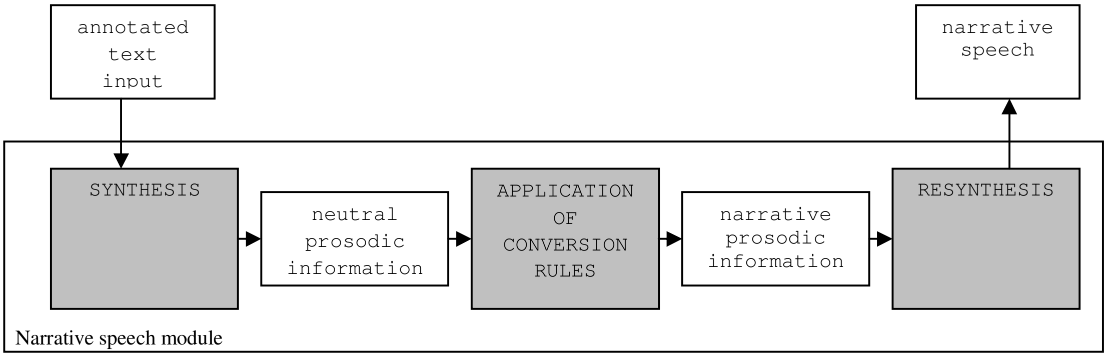

annotated text input
narrative speech
SYNTHESIS
neutral prosodic information
APPLICATION OF CONVERSION RULES
narrative prosodic information
RESYNTHESIS
Narrative speech module

Figure 9.1. Steps taken in automatic generation of narrative speech

The input text that is used has to be annotated in such a way that the module that is implemented knows how and where it should realise the prosodic functions narrative style and tension course (§2.5). Therefore some kind of markup language is needed, in which the input text can be written down so the positions where these functions must be applied can be indicated. We will use an existing markup language for this purpose and extend it with our prosodic functions. The selection and extension of this language is performed before the implementation takes place, because the implementation depends on what kind of input information it gets. This forms the first step of the implementation process and is described in paragraph 9.2.

After the first step is finalised we have obtained a strict data definition to which all of our input data should apply. Based on this definition we can now build the module that is responsible for the synthesis of narrative speech. The functions that the module has to perform are shaded grey in figure 9.1., white boxes represent data.

The first task of the module is to synthesise the text to obtain the basic prosodic information. The Fluency text-to-speech engine (§2.6.3) can return two types of data: a waveform synthesis of the input text or string of prosodic information which is nothing more than a list of phonemes followed by their duration and pitch values. So by retrieving this prosodic information after the first synthesis step, we obtain a preliminary neutral pronunciation of the input text in a form we can easily process.

The next step is to modify the prosodic data we got in the previous step such that the desired narrative functions are realised, resulting in prosodic output data in the same form as the input data. Now the central part of the implementation is reached, in which the conversion rules (chapter 5) are applied to the preliminary neutral pronunciation[17]. This is where the XML annotation gets involved, since the module needs to know in which positions of the speech data the conversion rules should be applied. After the rules have been applied to the complete preliminary pronunciation, we have obtained a series of phonemes with duration and pitch values which are ready to be resynthesised by Fluency. So we will use these values as input for the resynthesis process, in which Fluency generates a new pronunciation based on the narrative prosodic information. This time we will have Fluency perform a waveform synthesis, so the output we get is a speech signal. Since we have obtained our desired output this is where the module has finished its tasks.

In paragraph 9.3 we will discuss the implementation of the narrative speech module in more detail. Not all the steps we describe above will be discussed, because some are not relevant enough. The focus will be on the most important part of the implementation, the application of the conversion rules.

9.2 Markup language

9.2.1 Introduction

In this section we will take a closer look at some existing markup languages that are eligible for use in our implementation and select one that meets our requirements. One reason we use an existing markup language and not define one ourselves is that existing languages already offer a well-documented structure that is needed to define data at a prosodic level, so the existing language can be easily extended to meet our requirements. Another reason is that some standards support the definition of data from a wide perspective (for example description of gestures, facial expressions and speech together), which would make it fit the properties of the Virtual Storyteller project, in which various ways of embodied agent expression are involved.

Several markup languages exist that support annotation for text-to-speech purposes, some have a broader approach than others and for example they are aimed at providing a markup language for embodied conversational agents, including non-verbal style.

The languages that we will compare are the following:

- GESTYLE
- SABLE
- SSML (Speech Synthesis Markup Language)
- AML (Avatar Markup Language)

In the following paragraph (§9.2.2) we will discuss the properties of each of the languages, afterwards we will choose on of the languages to use in our implementation (§9.2.3).

In order to make the markup language we selected suitable for use in our implementation some changes have to be made to its structure, which is defined in a DTD. This structure change is described in paragraph 9.2.4.3.

[17] As already indicated in the text-to-speech engine selection (§5.55.5) Fluency doesn't offer the possibility to control the intensity of the speech at phoneme level, so we will not include the intensity conversion rule in the implementation.

9.2.2 Description of markup languages

9.2.2.1 GESTYLE

GESTYLE is aimed at representing meaningful behaviours of embodied conversational agents (ECA) with both non-verbal style and speech style explicitly given [13]. GESTYLE is an XML compliant language which can be used to define style and to instruct the ECA to express some meaning both verbally and non-verbally. GESTYLE acts both at a high and low level of description, for example by defining turn taking in a conversation but also defining specific hand gesture instructions.

GESTYLE uses a hierarchical approach to make the specification of instructions at different levels possible. With respect to speech style, GESTYLE is intended to use the same approach, so certain high level styles can be expressed in lower level styles (defining tag 'happy_speech' as a certain increase in pitch and speech rate, which are also tags in GESTYLE). The speech style can be defined by using both speech property modifiers (global pitch or speaking rate) and phoneme level modifiers (duration and pitch applying to individual phonemes). From the perspective of our implementation GESTYLE could be used as markup language to describe our data throughout the entire implementation process. Because of its hierarchical description possibilities it could first be used to describe the input data at a high prosodic level (accents and climaxes). Then at the point of the application of the conversion rules GESTYLE could be used to describe low level prosodic data (phonemes with pitch and duration)[18].

GESTYLE is still in an experimental phase of development, so no ready-to-use version is at our disposal.

9.2.2.2 SABLE

SABLE is an XML markup language aimed at providing a single standard for speech synthesis markup [24]. SABLE uses tags to specify emphasis, pitch, intensity and speech rate in a certain text, both in an absolute (by specifying the adjustment by an absolute value in the property's quantity) as in a relative way (by specifying the adjustment by a percentage or a descriptive term like 'small' or 'large'). There is no hierarchical approach with respect to tag structure in SABLE. The possibilities SABLE offers are relatively limited with respect to other languages and seems to be somewhat outdated, though SABLE is simple and directly usable.

9.2.2.3 SSML

Currently being a candidate recommendation by W3C [25] SSML (Speech Synthesis Markup Language) is on its way to become a standard in synthetic speech markup. SSML is an XML-based markup language for assisting the generation of synthetic speech [26]. It offers a way to control aspects of speech such as pronunciation, volume, pitch, rate, etc. The activity carried out in SABLE was used as the main starting point for defining of the requirements of SSML, so SSML goes on where SABLE stopped and is therefore a more complete standard.. In SSML text-to-speech properties and styles can be described at both prosodic and phoneme level.

[18] In this case the text-to-speech engine should support the processing of prosodic data based on GESTYLE.

9.2.2.4 AML

AML (Avatar Markup Language) is another high-level XML language aimed at synchronisation of speech, facial expressions and body gestures [27]. The language is primarily used for description of avatar animation. Most of the tags included in the language however aim at the definition of facial expressions and body gestures, especially handling timing issues. AML provides no tags for the annotation of expressive speech, only a global text-to-speech tag is included.

9.2.3 Markup language choice

Starting point in the selecting of an appropriate markup language is that we have to add markup tags to the language that makes the annotation of narrative speech possible.

It is obvious that each of the four described markup languages for speech synthesis we consider has its own advantages and disadvantages. Whichever language we choose to use in our implementation, we will always have to modify it in order to make the markup of narrative speech possible. So one straightforward criterion that will be used in the judgement of the markup languages is that the language should be easily extensible. So if any other synthetic speech related tags are already present in the language, a language is regarded easily extensible because the tags we use can be based on the existing tag structure.

Another criterion that is of importance is the nature of the language itself. A language that is specially aimed at use in embodied conversational agent environments is more desirable than a language that is solely aimed at markup of synthetic speech, because this better fits our project environment. If a language supports facial expressions, gestures and speech definition, the plot creator in the Virtual Storyteller can generate one plot in which all these communicative aspects of the storyteller are incorporated.

Though the AML language is interesting from the point of view of the Virtual Storyteller project because its aim is at support of embodied conversational agents, there is minimal support for the definition of synthetic speech in the language. Therefore this language will not be used.

SABLE provides much more possibilities for the definition of prosodic information in texts. But since SABLE was used as a starting point in the development of SSML, SSML is preferred above SABLE.

GESTYLE is the only language that combines both verbal and non-verbal style in a hierarchical approach, which are characteristics that are desirable for use in our Virtual Storyteller project. Because of this approach the language is easily extensible and it offers sufficient possibilities for text-to-speech purposes. Although GESTYLE meets our requirements best of all four languages, it is still a language in development, which means no employable version that can be used for the description of synthetic speech is available yet, so we can’t use it in our implementation.

The last language we considered is SSML. SSML already contains a set of prosodic tags that are at the same level as our narrative speech tags. Adding our tags to SSML will be relatively easy because the tag structure is already available. A disadvantage of SSML is that it is only aimed at speech synthesis annotation so no support for embodied conversational agents is included. Since GESTYLE is not usable yet and SSML is the second best language that meets our requirements, we will use SSML as markup language in our implementation.

9.2.4 Adding narrative speech tags to SSML

9.2.4.1 Narrative speech tags

The prosodic functions we use in our project are general narrative style and tension course. So we have to add tags to SSML that make the annotation of those functions possible in the input texts of our implementation.

The first tag we need is a tag that specifies what kind of speaking style should be used in the speech, normal speaking style or narrative style. If narrative style is used in the annotated text the words that have sentence accent should be marked (§3.5), so the text to speech engine knows to which syllables the manipulations should apply. Furthermore from the analysis it turns out that the annotator should have the possibility to increase the duration of certain accented syllables. We will create a tag that is used to annotate sentence accents, this tag will have an attribute in which can be specified whether the concerning syllable should be extended as well. This tag is different from the already existing *emphasis* tag in the fact that we don't want to allow specification of the emphasis strength (which is defined for the *emphasis* tag by attribute *level*).

For the realisation of the tension course function (§2.2), which done by the climaxes, some tags are needed as well. In the analysis we distinguish two kinds of climaxes, the sudden (§4.3.2) and the increasing climax (§4.3.3). Two tags are needed to indicate the start and end of a sudden climax, and three are needed for the increasing climax: the first indicates the start of the climax, the second indicates the top, and the third marks the end of the climax. We will define only one climax tag that can be used for both types of climax; as a result we need to use an attribute inside the climax tag to indicate which kind of climax is used. An XML element always starts with an opening tag and ends with an ending tag, between the tags is the data that belongs to the element. So if we define a climax element we obtain two tags that can be used to indicate beginning and end of the climax. We do need an extra tag though to indicate the top of an increasing climax. This tag will of course only be necessary in the increasing climax and can be omitted in the sudden climax.

Summarising we need the following tags:

- speaking style tag with a type attribute
- sentence accent tag with an extension attribute
- climax tag with a type attribute
- climax top tag

9.2.4.2 The SSML DTD

First we will take a short look at the DTD that is used to define SSML, of which the main tag structure will be described here. To clarify the data that is described in a DTD it often helps to give a sample of XML data that complies with the DTD:

```
<speak ... >
  <p>
    <s>You have 4 new messages.</s>
    <s>The first is from Stephanie Williams and arrived at 3:45pm.</s>
    <s>The subject is <prosody rate="-20%">ski trip</prosody></s>
```

```
6:   </p>
7: </speak>
```

Each text that is to be synthesised starts with the `<speak>` tag which is also the root tag of the document. The text in the XML must be formatted by using paragraphs and sentence separation tags (tags `<p>` and `<s>`) . In line 5 we see that the text is interrupted by a `<prosody>` tag. In this case the attribute `rate` specifies that the speech rate during the words "ski trip" should be lowered by 20%.

So in general can be said that each text inside the <speak> tags must be structured in paragraphs and sentences, and that inside sentences prosodic tags can be used. Below is a part of the DTD that is used to describe the XML data[19]. We have stripped this DTD so only the tags that were used in the XML sample are visible ('…' means that something was removed from this tag definition).

```
<!ENTITY % duration "CDATA">
<!ENTITY % structure " p | s">
<!ENTITY % sentence-elements " prosody | ... ">
<!ENTITY % allowed-within-sentence " ... | %sentence-elements; ">

<!ELEMENT speak (%allowed-within-sentence; | %structure; | ... )*>
<!ATTLIST speak
    ...
>

<!ELEMENT p (%allowed-within-sentence; | s)*>
<!ELEMENT s (%allowed-within-sentence;)*>

<!ELEMENT prosody (%allowed-within-sentence; | %structure;)*>
<!ATTLIST prosody
    pitch CDATA #IMPLIED
    contour CDATA #IMPLIED
    range CDATA #IMPLIED
    rate CDATA #IMPLIED
    duration %duration; #IMPLIED
    volume CDATA #IMPLIED
>
```

The structure of this DTD is quite simple. Starting at the root node <speak>, this tag may contain a *structure* entity, which a *paragraph* or *sentence*, or it may contain an *allowed-within-sentence* entity. Among others this can be an entity *sentence-elements*, which can be a *prosody* tag. Looking at element prosody, we see that inside a *prosody* tag once again an entity *allowed-within-sentence* and entity *structure* are allowed, meaning that the *prosody* tag can contain data recursively. The *prosody* tag has possible attributes *pitch, contour, range, rate, duration* and *volume*.

9.2.4.3 Adaptation of the DTD

In paragraph 9.2.4.1 we have determined that the following tags are needed:

- speaking style tag with a type attribute

[19] It is assumed that the reader is familiar with DTDs and so this will not be explained here

- sentence accent tag with an extension attribute
- climax tag with a type attribute
- climax top tag

The first thing that will be changed in the DTD is that the global descriptive tag which indicates what kind of speaking style is to be used in the speech is added. The value of this tag determines whether the text-to-speech engine will use its normal pronunciation for the text (in fact ignoring all narrative tags), or use narrative style to pronounce the text. For this purpose we add following tag to the DTD:

```
<! ELEMENT style EMPTY>
<!ATTLIST style
     type (normal | narrative) “normal”
>
```

Furthermore the *speak* element is updated, because now inside *speak* tags the *style* tag is allowed to be used:

```
<!ELEMENT speak (%allowed-within-sentence; | %structure; | style | ... )*>
```

The *style* element has one attribute, which is the *type* of style that should be used in the speech (*normal* or *narrative*). The default value of type is *normal.*

The addition of the sentence accent tags goes in a similar way. We add the following tags:

```
1:  <!ENTITY % sentence-elements " prosody | sentence_accent | climax | ... ">
2:
3:  <!ELEMENT sentence_accent (#PCDATA)>
4:  <!ATTLIST sentence_accent
5:     extend (yes | no) #REQUIRED
6:  >
7:
8:  <!ELEMENT climax (#PCDATA | climax_top )*>
9:  <!ATTLIST climax
10:    type (immediate | increasing) #REQUIRED
11: >
12: <!ELEMENT climax_top EMPTY>
```

The entity *sentence-elements* on line 1 now contains two new elements: *sentence_accent* and *climax*, meaning these tags can now be used inside a sentence.

The *sentence_accent* element is defined on line 3. The element only allows text within its own tags, so no nested other tags are allowed. The *sentence_accent* element has attribute *extend* which specifies whether the syllable that is inside the *sentence_accent* tags should be increased in duration.

On line 8 the climax is defined. Between the *climax* tags a *climax_top* and plain text is allowed. On line 12 can be seen that this is an empty tag, its only function is to mark the position in which the climax has its top. The *climax* tag also has an attribute *type*, defining the nature of the climax: *immediate* or *increasing*.

The following sample of validated XML data is an example of text that contains all of the tags that are necessary for annotating narrative speech:

```
<speak ... >
  <p>
    <s>
      Die baard maakte hem <sentence_accent extend="yes">zo</sentence_accent>
      afschuwelijk lelijk dat <sentence_accent extend="no">ie</sentence_accent>dereen
      op de loop ging zo<sentence_accent extend="no">dra</sentence_accent> hij in de
      buurt kwam.
    </s>
    <s>
      Hij wilde zich omkeren <climax type="imediate">en toen</climax> klonk er
      <sentence_accent extend="no">plot</sentence_accent>seling een harde knal
    </s>
    <s>
      Blauwbaard hief het grote mes op, <climax type="increasing">hij wilde toesteken
      en <climax_top/>toen werd er hevig op de poort geklopt.</climax>
    </s>
  </p>
</speak>
```

We have chosen to explicitly mark sentence accent syllables and not sentence accent words. This is not the most elegant approach because it would be more proper to annotate the entire word and have the text-to-speech engine determine the correct syllables that the accent must be applied to. The problem here is that in its prosodic output Fluency doesn’t always return the accented syllables of words that are synthesised.

The prosodic output of Fluency always contains a transcription string of phonemes in which word accents are denoted by an apostrophe (‘’’) and pitch accents by an asterisk (‘*’). If for example we synthesise the sentence “Hij komt altijd te laat binnen.” we obtain the following transcription string:

```
HEi k'Omt Al-tEit t@ l'at bI-n@n0
```

If we want to manipulate the sentence accents of this sentence in order to produce narrative speech, it is a possibility that stress must be put on the first syllable of “altijd” and the first syllable of “binnen”. But Fluency doesn’t supply any information about which of the syllables of those words has word accent. Although in a lot of cases Fluency does supply the word accents of all words (of more than one syllable), it also frequently happens that the word accent of one or two words of a sentence is not supplied. This forces us to annotate the sentence accents on syllable level, so the specific syllable which the accent manipulation must be applied to is known to the module that applies to manipulations. We will therefore not base the manipulations on the accents that Fluency gives in its prosodic output.

Besides in storytelling it is possible that relatively many words must be accented in a sentence, because the storyteller wants to emphasize some extra words. If we would use Fluency for the detection of sentence accents this would restrict us to the number of accents that Fluency supplies, which may not be enough for the purpose of storytelling.

9.3 Implementation of narrative text-to-speech module

9.3.1 Introduction

We start the description of the implementation by explaining the general architecture of the module that is implemented (§9.3.2). The rest of this paragraph describes the process in which the conversion rules for narrative speech (chapter 5) are applied to the synthesised annotated input (§9.3.3).

The description of the implementation will start at the highest level in which we have the annotated input text, and step by step we will follow the process, ending with the description of the manipulation of phonemes.

Not all aspects of the implementation will be described in equal detail. The general process that is executed and all the steps involved in it will be explained at informal level. Because for the manipulation of the two prosodic functions narrative style and tension course almost similar processes are gone through, we will focus at only one of these in the description of the implementation, namely narrative style. In the implementation of the tension course function roughly the same steps are taken.

Because the manipulation of the narrative style and tension course involves similar processes, we will focus on only one of these in the description of the implementation, namely narrative style. In the implementation of the tension course function roughly the same steps are taken.

During the explanation of the implementation references will be made to certain functions. Including the code of the functions in this report is undesirable because of its extent, therefore the code can be found on the World Wide Web [31].

9.3.2 General architecture

The module is implemented in Java, which is the common implementation programming language for the Virtual Storyteller project. The process that has to be executed is completely sequential, because a certain input is processed step by step in the end resulting in the output of the module. Because of this sequential character of the process we have implemented only one class, the *Storyteller* class. This class consists of a set of functions of which each is responsible for a certain step in the sequential process.

As described in the introduction (§9.1) during the execution process of the module the Fluency TTS engine is needed twice. To be able to make use of the Fluency TTS engine the Fluency DLL is used. A Java wrapper developed at our HMI group is used in order to import the C++ Fluency functions from the DLL in our module. By importing the class of the Java wrapper we can use the Fluency functions in our S*toryteller* class.

In fact only two functions implemented by the Java wrapper are of importance. The first is a function that sets the global speech rate of the text-to-speech engine. The other function is the function that performs the actual speech synthesis.

9.3.3 Application of conversion rules

9.3.3.1 Introduction

To be able to apply the conversion rules to the preliminary neutral pronunciation of the input text, we first have two know two things. First we need to know the representation format of the prosodic information that we have to manipulate, which is explained in paragraph 9.3.3.2.
Besides we have to find out to which phonemes of the prosodic information the manipulations must be applied. The first step that is involved here is to find the tag indicating the prosodic functions in the XML input, which can be done by parsing the XML (§9.3.3.3). After these have been found their positions should be looked up in the prosodic information, so manipulations can be applied in the correct place. This mapping process is explained in paragraph 9.3.3.4.
After the correct positions in the prosodic information are located, the data can be manipulated. This process is described in the last paragraph of this section (§9.3.3.5).

9.3.3.2 Prosodic information format

As explained in the introduction (§9.1), a preliminary pronunciation of the plain input text is obtained by synthesising the input text without taking into account the narrative speech notation. Therefore we first have to strip all tags from the XML input data, resulting in a plain text string consisting of the sentences that have to be synthesised (the annotation is removed only temporarily to get the preliminary pronunciation and is stored for later use in the application of the conversion rules).
After we have obtained this plain text string we use it as input for the Fluency engine, ordering the engine not to return waveform data but the prosodic information. Fluency will synthesise the string, returning the prosodic information in the form of a string of phonemes with accompanying duration and pitch values. A sample of the prosodic information that is returned is given here:

```
1: h 112
2: I: 151 50 75
3: R 75
4: l 75
5: @ 47 20 71 70 61
6: k 131
7: @ 55 80 70
8: _ 11 50 65
```

Each line starts with a phoneme (using SAMPA notation) followed by the duration of that phoneme in milliseconds. A pause is denoted by an underscore ('_') character[20]. If a certain phoneme is a vowel, it is possible that the phoneme has pitch value(s). A pitch value always starts with a percentage followed by a pitch value. This percentage indicates at which point during the vowel the pitch should be applied. The pitch value that follows the percentage is expressed in Hertz. If for example we consider the 'schwa' on line 5, we see that this vowel should last for 47

[20] A pause has a pitch value in Fluency's prosodic format. It is for example used to define the start and end pitch of a sentence.

milliseconds. After 20% of the duration is elapsed, the pitch should be changed to 71 Hz, after 70% of the duration is elapsed, the pitch is changed to 61Hz.

9.3.3.3 XML Parsing

Java contains standard packages with functions for the parsing of XML data. The first thing to do is read the XML input text into an XML document object. XML can be parsed by using a function that returns a list of certain XML elements that can be specified by tag name.

Any tags we use in our input text for the indication of sentence accents and climax must start and end in the same sentence. For the sentence accent this is obvious, because it acts on syllable level, but although a climax is allowed to contain multiple words, a requirement is that the climax is not spread over more than one sentence. Besides, from the analysis perspective (§4.3.1) this is not desirable, since climaxes only apply to one sentence.

This brings us to the approach to process the XML data by sentence. Each sentence is a node in the node tree of the XML document (using DOM, [28]), so we will one by one select the sentences and check if they contain a sentence accent tag or climax tag. If one of these tags is found, a function is called that will find the positions in the prosodic information that must be manipulated (§9.3.3.4, function `applySentenceAccents`) and do the actual manipulation in the sentence in which the accent or climax tag was found (§9.3.3.5). After all sentences have been processed in this was the parsing is finished.

9.3.3.4 Mapping

This paragraph describes the procedure after a sentence accent or climax is found in a sentence of the XML input. The description is based on the sentence accent, a similar approach is followed for the climax.

In the case of a sentence accent, at this point it is known that a certain syllable in the sentence is to be manipulated. The manipulation must take place in the prosodic information of the preliminary pronunciation, which is a string of phonemes plus their prosodic properties (from here on referred to as the *phoneme string*). So the syllable that is to be manipulated must be looked up in the phoneme string and must be manipulated afterwards. In order to find the relevant syllable in the phoneme string, we must make a mapping from the syllable (which is written down in plain text) to the phoneme string. This is an operation that involves several sub steps, that will be explained here.

The biggest problem here is that we only know what the position and the characters of the syllable are (they are between the *sentence_accent* tags) in the plain input sentence, and that in the phoneme string a different notation is used to denote the phonemes (SAMPA notation). So for each character in the syllable a mapping must be made to its possible phonetic appearances. Before we can do this, we have to be certain that if a match is found in the phonetic string, this is the correct syllable, because it is conceivable that a certain syllable appears more than once in the phonetic string. Consider for example the following annotated input and the accompanying sample of the phoneme string:

```
De <sentence_accent extend=”no”>koe</sentence_accent>koek had een slechte dag.
```

```
1: k 140
2: u 103 50 126
3: k 112
4: u 71 20 101
5: k 140
```

If we would just map the ‘koe’ syllable to the phoneme string it may result in selection of the wrong phonemes, because the phonemes on line 3 and 4 may be selected. To avoid this problem we will include the neighbours of the syllable in the mapping.

Here a consideration must be made about the number of neighbour syllables or words we want to include in our mapping from syllable to phonemes. The first thing we can do is only include the neighbour syllables of the sentence accent syllable in the mapping. This means the complete word of the sentence accent syllable will be used for the mapping. This is still not a safe choice either, because it frequently happens that the same word occurs more than once in a sentence. As a result we have to include some of the neighbouring words in the mapping as well. By including the word before and the word after the word in which the accented syllable occurs, it is almost certain the correct word is selected.

There is still a small chance that a wrong mapping occurs though if the same sentence contains two sequences of exactly the same words (an example for which this is the case is ‘als mollen mollen mollen, mollen mollen mollen’). To solve this it was also possible to include more neighbour words in the search. Disadvantage of this is that this brings along more processing. This is where the consideration has to be made, in which there is a trade off between the chance of correctness of the results and the amount of processing involved in the mapping. Because there is a small chance that a certain word sequence of three words is repeated in a sentence we will use word trigrams in our mapping procedure.

As said before, the mapping involves several sub steps that must be carried out:

- Find the index word of the accented syllable in its sentence
- Find the index and length of the accented syllable in the prosody string
- Use of regular expressions for character to phoneme mapping

These steps will be explained here one by one.

Find the index word of the accented syllable in its sentence

The goal of the use of word trigrams is to find the exact position of the accented syllable in the phoneme string. In the phoneme string words are separated by a certain separating character, so if we find out the index of the word that has to be manipulated, the corresponding phonemes can be selected. So first thing to do is to find out in which word of the sentence the sentence accent is annotated. This seems a simple operation which can be done by just counting the number of words before the annotated sentence accent word, but in the XML document this isn’t possible because the text of a sentence is split up in nodes. So if a sentence accent tag is found while parsing the XML, we don’t know exactly how many words of the sentence are before that tag, and if another tag of any kind is preceding it. Of course we can find this out by traversing back in the node tree of the sentence, and counting all words in preceding nodes. But it is easier to use another approach.

The approach we use here (function `applySentenceAccents`) is to first strip all XML tags from the input document, leaving a sequence of sentences. We can split this sentence sequence based on the dot (‘.’) that ends each sentence. Now on one side there is a certain sentence containing plain text and on the other side the word trigram belonging to the accented syllable of the same sentence. The next step is to try to match the word trigram to the plain text sentence, giving us the exact word index of the accented syllable’s word.

Find the index and length of the accented syllable in the phoneme string

Starting point here is that we know the sentence position of the word that has to be manipulated, so we can select its corresponding phoneme word from the phoneme string. Since the manipulation of the sentence accent takes place on syllable level, we have to locate the accented syllable in the phoneme word. This would be an easy task if the phoneme string would contain syllable separation marks, but this is not the case, as can be seen in the following phoneme example:

```
1: k 140
2: u 103 50 126
3: k 112
4: u 71 20 101
5: k 140
```

But there is another solution for this problem. In addition to the prosodic information of a synthesised text, Fluency also returns a *phonetic transcription string* of the synthesised text (from here on referred to as the *transcription string*). This string is a series of the phonemes of the sentence, including accents (‘’’ and ‘*’), sentence separators (‘ ’) and syllable separators (‘^’). For example the transcription string of the sentence “De koekoek had een slechte dag.” is:

```
^D@ ^k*u-k'uk ^H'At ^@n ^sl*Ex-t@ ^d*Ax
```

We can select the correct word in the transcription by using the word index (function `manipulateSyllable`). From the plain input sentence we already know whether the accented syllable has any preceding or following syllables. If any of these two options is not the case, the selection of the accented syllable in the transcription string is easy because the corresponding syllable is at the beginning or the end of the word, enabling us to select the similar syllable from the transcription string. If the accented syllable has both preceding and following syllables, we have to find it in another way (function `findSyllablePosition`).

The approach here is that we will use regular expressions (function `createRegEx`) for the mapping of alphabet characters to phonetic symbols. The mapping by regular expressions itself will be explained in the following section, here it is enough to know that given a certain sequence of alphabet characters a regular expression mapping returns the corresponding sequence of phonemes. Assume we want to map the following annotated word to its corresponding transcription string:

```
ver<sentence_accent extend="no">an</sentence_accent>deringen
^v@r-*An-d@-r'I-N@n0
```

We can do this as follows. First we get the syllables preceding the accented syllable from the XML input (‘ver’). We use these in the regular expression mapping, giving us the phonetic transcription of the preceding syllables (“`v@r`”). Then we remove this sequence of phonemes from the transcription string. The same procedure can be followed for the following syllables of the accented syllable (‘deringen’). Regular expression mapping yields the phonetic transcription (‘`d@rIN@n0`’), which can then be removed from the transcription string. These operations leave the phonetic transcription of the syllable that is looked for (‘`An`’), including its position and length in the transcription string[21].

Now it seems that since we found the position and length of the accented syllable in the transcription string we can use the same position and length in the phoneme string and go on manipulate it. This is not correct because one more step must be carried out. It turns out that the phonemes in the transcription string aren’t always similar to the phonemes in the phoneme string. This is because the transcription string is nothing more than the phonetic representation of the individual words taken from a lexicon that Fluency consults during the synthesis process. The phonetic string on the other hand, is the phonetic representation after Fluency has applied certain phonetic rules that are important in speech synthesis, such as rules for assimilation and reduction [1]. One example of a difference between transcription and phoneme string is that if one word ends with a certain phoneme and the next word starts with the same phoneme, the two phonemes are taken together. So in the transcription string both phonemes are given, but the phoneme string contains only one occurrence of the phoneme.

To avoid selecting the wrong phonemes because of the possible differences between transcription and phoneme string, we will map the transcription string to the phoneme string (function `getMapping`). Now that the position and length of the syllable in the transcription string and the mapping of the transcription string to the phoneme string are known, the position of the accented syllable in the phoneme string is found and we can start to manipulate its corresponding duration and pitch values. Before we will describe this (§9.3.3.5) we will first describe the regular expression mapping.

Use of regular expressions for character to phoneme mapping

A regular expression[22] is used to map alphabet characters on one side to phonemes on the other side (function `createRegEx`). The reason regular expressions are used for the mapping is that a certain alphabetic character can be realised by several different phonemes depending on its position in a word. To map an alphabetic word to a phonetic word, a regular expression is created that contains all possible realisations of the characters in the alphabetic word. This expression can then be matched against the phonetic word, resulting in a character to phoneme mapping.

[21] This approach guarantees that the correct syllable is returned. If we would not use neighbour syllables in the mapping we might end up with the wrong syllable because the same syllable can occur several times in a word.

[22] The reader is assumed to be familiar with regular expressions

An example of regular expression construction is the following. Character ‘g’ can be realised phonetically by ‘G’ or ‘Z’ (‘**g**oed’ and ‘baga**g**e’). Besides the ‘g’ character can occur in the alphabetic sequence ‘ng’, which is realised phonetically by ‘N’ (‘ba**ng**’). If a ‘g’ is the current character at regular expression creation time, the following pattern is added to the expression:

```
((G|Z)?)
```

This means the ‘g’ can either be realised by ‘G’ or ‘Z’ or no phonetic occurrence is necessary (‘?’). The latter is the case when there really is an occurrence of ‘ng’. If this is so at the moment the ‘n’ was processed, the phonetic ‘N’ was already appended to the regular expression. It is incorrect to do this again when subsequently the ‘g’ is processed, so nothing (realised by ‘?’) should be added to the expression in this case.

The same process that was illustrated in the example is performed for all characters in the alphabetic word. After the regular expression is completed it is matched to the phonetic word, resulting in the character to phoneme mapping.

9.3.3.5 Manipulation of phonemes

Before we go on describing the manipulation of phonemes one remark must be made. During the formulation of the conversion rules (chapter 5) we observed that a lot of variation occurs among the values of certain constants that appear in the conversion rules. To formulate a deterministic set of rules in the constant evaluation the best values for those constants were derived. For the rest of this project those rules were used as a starting point (including the implementation), although we already stated in the conversion rule formulation that there’s still the possibility that a stochastic model is needed to vary the constant values within a certain range. During the implementation of the module we performed a small test to see whether a stochastic approach results in more variation. The normal constant value that is used for the desired maximum pitch increase in accented syllables is 40 Hz, in our small test we randomly generated a pitch value in the range [30, 40] Hz for this constant. The test resulted in fragments that at first sight sound more natural because of their variation. To find out for which constants this variation provides an important contribution to the naturalness of the speech is a question that will not be answered here. It would require an extra evaluation; therefore we will include this in the recommendations (chapter 12).

Two kinds of manipulations can take place: the manipulation of duration (function `adaptDuration`) or the manipulation of pitch of a phoneme (function `adaptSyllableAccentPitch`. The manipulation of the pitch is the more interesting of the two; therefore it will be described here.

The manipulation of the pitch is based on the pitch rule for accented syllables that is given in the conversion rules (§5.2.1). We will repeat this rule here, but we have substituted values of constants m_1 and m_2 as determined in the constant evaluation (§7.4). A single syllable pitch manipulation for a syllable in domain $[t_1, t_2]$ is performed by the following formula:

$$y'(t) = \begin{cases} y(t).(1 + (\sin((((t - t_1)/(t_2 - t_1))0{,}5pi) + 0{,}25pi)/n)) & ,if \quad t \in [t_1, t_2] \\ y(t) & ,else \end{cases}$$

The following variables and constants are used:

`y'`	manipulated pitch values
`n`	constant determining the degree of adaptation

From the constant evaluation it turns out that the desired maximum pitch increase constant should be 40 Hz. Consequently the following formula is used to calculate the value of `n`:

$$n = avg\_pitch / 40$$

The data that is needed in order to execute the rule will be described first.
Because each phoneme in the phoneme string has its own duration value we don't have a time domain which starts somewhere before the syllable, so the first phoneme of the syllable is assigned time value $t = 0$. In the formula the total duration of the syllable is used as well (t_2-t_1), in the implementation this is calculated by adding the durations of all phonemes in the syllable. For each phoneme in the syllable the formula is executed, after the execution variable t is incremented by the phoneme's duration. In this way we guarantee that the right value of the sine function is used as base for the pitch adaptation.
Another value that is needed is the pitch value of the phoneme that is manipulated. Of course this value can be directly read from the phoneme string, but it is possible that a vowel doesn't have a pitch value specified. In this case the first pitch value that occurs in its preceding phonemes is used.
The last value that is needed in the calculation is the average pitch value, which is used to calculate `n`. A problem here is that it is hard to find out what this value is. Of course it would be possible to process all pitch values in the phoneme string and then calculate the average, but this involves a lot of processing. Another possibility is to define the average pitch as a constant, based on the voice that is used by Fluency in the speech synthesis. This is the safest approach and therefore it will be used in the implementation.
Now that all input values are calculated the pitch of all phonemes of the accented syllable can be manipulated based on the formula. After all pitch and duration manipulations have taken place, the manipulated phoneme string is returned.

10 Implementation evaluation

10.1 Introduction

The implementation evaluation is the final project phase and is aimed at determining the added value of the application of the conversion rules with respect to the quality of storytelling. This seems similar to the evaluation that was carried out before, the conversion rule evaluation (chapter 8). The difference here is that the narrative speech module has been implemented at this stage, which was not the case at the moment of the conversion rule evaluation. The fragments that were created for the purpose of the conversion rule evaluation were all created by first having Fluency create a neutral pronunciation of the fragment, and subsequently using *Praat* to manipulate its acoustic features (§6.3) and create a resynthesis of the fragment. The fragments that are created in the implementation evaluation on the other, are created by manipulation by our module and resynthesis by Fluency (§9.1).

Another difference between the two evaluations is that the conversion rule evaluation was an evaluation with a small group of participants, while in the implementation evaluation a larger group of participants will be used.

The set-up of this evaluation is exactly equal to that of the conversion rule evaluation. The only difference is the stimuli that are created and the size of the participant group. For this reason the choice of questions will not be explained again here, for this is already described in paragraph 8.2. A short overview of the questions and stimuli will be given though (§10.2). Afterwards a hypothesis will be formulated (§10.3) followed by a description and discussion of the result of the evaluation (§10.4). The chapter ends with conclusions of the implementation evaluation (§10.5).

10.2 Questions and stimuli

During the conversion rule evaluation it turned out that sometimes participants had the tendency to be looking for a certain phenomenon in the fragments (§8.4.1), because they expected that every fragment was manipulated in a certain way. To avoid influence of the results by this kind of bias in the implementation evaluation, we added a remark to the introduction text of the experiment saying that participants shouldn’t expect tension and storytelling aspects in all fragments (appendix F).

The series of fragments will be evaluated based on the 5-scale method by Likert [11]. We want to evaluate the speech fragments based on certain judgement criteria. The criterions to base the evaluation on are the quality of the storytelling, the naturalness of the speaker and how tense the fragment sounds according to the participant. Based on this we will accompany each fragment with three questions and ancillary answers (table 10.1):

Question	Answer range
“Hoe goed vind je de spreker voorlezen?” (“How do you judge the quality of storytelling of this speaker?”)	• 1 = ‘zeer slecht (‘very bad’) • … • 5 = ‘uitstekend’ (‘excellent’)

"Hoe natuurlijk klinkt de uitgesproken tekst?" ("How do you judge the naturalness of the fragment?")	• 1 = 'zeer onnatuurlijk' ('very unnatural') • … • 5 = 'zeer natuurlijk' ('very natural')
"Hoe spannend vind je het fragment?" ("How tense do you think the fragment sounds?")	• 1 = 'niet spannend' ('not tense') • … • 5 = 'heel spannend' ('very tense')

Table 10.1. Conversion rule evaluation questions

The total set of stimuli consists of *sixteen* stimuli, of which eight are neutrally spoken unique text fragments and eight are the same unique text fragments spoken in narrative style or climax. These last eight stimuli can be divided in five containing narrative style only, and three containing both narrative style and climaxes. We will create two test sets across which we divide the sixteen stimuli; the first test set is to be evaluated by one half of the participant group, the other set by the other half. We will divide the stimuli in such a way, that each unique text fragment is presented in neutral form to one group, and the same fragment in manipulated form to the other group. The following schema summarises the above (figure 10.1):

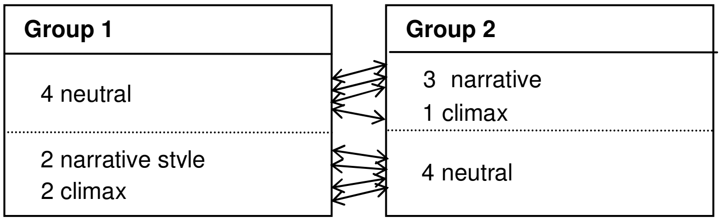


Figure 10.1 fragment division

The sixteen stimuli will be divided over two groups of participants of each 10 participants. The eight stimuli of each group will be provided in random order so that it is not possible to guess the nature of the fragment (narrative or climax) based on the order of the fragments.
The list of stimuli is provided in full detail in appendix E.

10.3 Hypothesis

The null hypothesis is that for a certain fragment, the manipulated version is not rated significantly better than the neutral version with respect to narrative quality, naturalness and tension display. So the average rating of the three judgement aspects of both versions is expected to be equal. We will reject this hypothesis if there is valid proof based on mean judgement comparisons that the two are not equal.

10.4 Results

10.4.1 Introduction

The conclusions of the conversion rule evaluation (§8.5) state that:

1. In some cases addition of narrative style increases the appreciation of storytelling.
2. Naturalness of narrative speech is not experienced higher than that of neutral speech.
3. More tension is perceived in the case of climax presence.

Only the last conclusion was fully based on statistical proof of significance. Regarding significance, the results of the implementation evaluation show similar characteristics, so these results will be discussed in a qualitative way again. So although the method of fragment creation differs for the two evaluations, this doesn't give us such degree of increase in speech quality that we now have results in which the application of the conversion rules yields significant differences in judgement among the neutral and narrative speech fragments. This is because although Fluency is now used for the resynthesis of narrative speech after the conversion rules have been applied (§9.1), Fluency still uses signal processing methods (MBROLA, §2.6.2) which inevitably incur distortion if relatively large pitch and duration adaptations take place (§7.2 and [12]).
It is possible though that the increase in size of the participant group is of influence on the significance of the results. In the conversion rule evaluation the group was small, resulting in a small data set for statistical analysis. Although we expect that the judgements of the fragments in the implementation evaluation will not differ much from the judgements in the conversion rule evaluation, the increase in group size may result in more statistically significant differences in results. So in the description of the results of the implementation evaluation besides the standard statistical measures mean, standard deviation and range we will again include t-test and Mann-Whitney test results, which are used to see if any significant difference between the two populations exists.

10.4.2 Statistical results and discussion

This section contains the results and interpretation of the implementation evaluation. We will successively discuss the results of the three questions that were asked in the evaluation. The following tables show the mean, standard deviation and range of the answer values of the first question ("How do you judge the quality of storytelling of this speaker?"). Table 10.2a shows these statistics for non-manipulated fragments (indicated by "**a**"), table 10.2b for manipulated fragments (indicated with "**b**"). Fragments *1* until *5* only contain narrative style; *6* until *8* contain both narrative style and climax. We also applied the t-test and Mann-Whitney method to each couple of *a/b*-fragments. The significance with which can be said that the two series of results are statistically different is listed in table 10.2c.

Fragment	**1a**	**2a**	**3a**	**4a**	**5a**	**6a**	**7a**	**8a**
mean	3,0	3,1	3,1	3,0	2,5	3,1	3,1	3,0
standard deviation	0,9	1,0	1,3	1,2	1,1	0,9	1,1	0,8
range	3	3	4	4	3	3	4	2

Fragment	**1b**	**2b**	**3b**	**4b**	**5b**	**6b**	**7b**	**8b**
mean	3,9	3,5	3,3	3,6	3,2	3,6	3,5	2,8
standard deviation	1,0	1,0	0,9	0,8	0,9	0,7	0,7	0,8
range	3	3	3	3	3	2	2	4

Fragment	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**
t-test significance	0,05	0,38	0,70	0,20	0,14	0,18	0,35	0,58
Mann-Whitney significance	0,05	0,55	0,66	0,19	0,17	0,12	0,38	0,60

Table 10.2 a,b,c. Statistics for question "How do you judge the quality of storytelling of this speaker?"

Looking at the absolute mean answer values, it can be seen that all neutral fragments are considered to be of average storytelling quality, and that all manipulated fragments are considered to be of above average storytelling quality (on the Likert scale [11] a judgement value of 3 is the average).

Comparing the mean answer values of neutral fragments to that of manipulated fragments, it can be seen that all but one manipulated fragment (fragment 8) have higher mean answer values than their neutral counterparts. The mean difference is the largest for fragments 1, 4 and 5 (difference of respectively 0,9, 0,6 and 0,7). Based on the mean values we can say that in almost all cases manipulated fragments are considered to be of better storytelling quality. The neutral fragments on average have higher standard deviations than the manipulated fragments (total average of 1,04 against 0,85), meaning there is more unanimity in the judgement of manipulated fragments, which is also visible in the smaller average ranges of answers (total average of 3,3 against 2,9).

If we look at the significance values of the mean differences calculated by t-test and Mann-Whitney, there only turns out to be statistically acceptable significance in the case of the first fragment (fragment 1), assuming we employ the requirement that significance is acceptable if equal or below 0,05. Looking at the rest of the fragments it can be seen that fragment 4, 5 and 6 have the lowest significance values.

The following table (table 10.3 a,b,c) successively shows the statistics for the question "How do you judge the naturalness of the fragment":

Fragment	**1a**	**2a**	**3a**	**4a**	**5a**	**6a**	**7a**	**8a**
mean	2,6	3,3	2,6	2,6	2,5	2,5	3,1	3,1
standard deviation	1,0	1,1	1,1	1,2	0,8	1,1	1,0	0,6
range	3	3	3	3	3	3	3	2

Fragment	**1b**	**2b**	**3b**	**4b**	**5b**	**6b**	**7b**	**8b**
mean	3,7	3,2	2,8	3,3	2,3	3,2	3,5	2,9
standard deviation	1,3	1,0	1,0	0,9	0,8	0,8	0,8	0,7
range	4	3	3	3	3	2	3	2

Table 10.3 a,b. Statistics for question "How do you judge the naturalness of this fragment?"

Fragment	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**
t-test significance	0,05	0,83	0,68	0,16	0,60	0,12	0,35	0,51
Mann-Whitney significance	0,04	0,66	0,87	0,20	0,51	0,13	0,44	0,40

Table 10.3 c. Statistics for question "How do you judge the naturalness of this fragment?"

The judgement of naturalness shows more divergence than the judgement of storytelling quality. Looking at the absolute mean answer values we see that there's a lot of divergence for both neutral and manipulated fragments, so no general remark can be made here. One phenomenon that was observed in the free responses that were given by the participants in conversion rule evaluation (§8.4.1) is observed here again. From some remarks it turns out that the positioning of the sentence accents in the fragments is a source of low naturalness judgement, because some participants perceive the accents as being in the wrong place, although they were deducted from original storyteller speech.

When comparing the mean answer values of neutral and manipulated fragments, in five cases the manipulated fragment is considered more natural than the neutral one, and in three cases the opposite is the case. Of those five cases in which the naturalness of the manipulated fragment is judged higher, three cases show a relatively high difference. Those cases are fragment 1, 4 and 6 (difference of respectively 1,1, 0,7 and 0,7). In the three cases in which the neutral fragment was considered more natural than the manipulated fragment, the difference in mean judgement is low (about 0,2). The standard deviation and range of the fragment judgements show little difference between neutral and manipulated fragments.

Once again only fragment 1 has a statistical difference of means that has a significance level below or equal 0,05 for both the t-test and the Mann Whitney test. Fragment 4 and 6 are the fragments that closest approach this boundary.

The last series of tables (table 10.4 a,b,c) shows the statistics for the question "How tense do you experience the fragment?":

Fragment	**1a**	**2a**	**3a**	**4a**	**5a**	**6a**	**7a**	**8a**
mean	2,1	2,5	2,5	2,1	1,8	2,3	2,7	2,4
standard deviation	0,9	1,2	1,1	1,2	0,8	0,8	1,2	1,3
range	3	3	3	3	2	3	3	3

Fragment	**1b**	**2b**	**3b**	**4b**	**5b**	**6b**	**7b**	**8b**
mean	3,7	3,1	2,8	3,0	2,2	3,6	3,4	4,0
standard deviation	1,1	0,9	0,8	1,1	0,8	0,7	1,0	0,7
range	3	3	2	4	2	2	3	2

Fragment	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**
t-test significance	0,00	0,21	0,49	0,09	0,27	0,00	0,16	0,00
Mann-Whitney significance	0,00	0,33	0,53	0,09	0,26	0,00	0,14	0,00

Table 10.4 a,b,c. Statistics for question "How tense do you experience the fragment?"

Looking at the absolute mean values, it is clear that the tension experienced in all neutral fragments is low (mean is below 3 for all fragments). Although the climaxes are primarily aimed at increasing the communication of tension, it seems that the addition of narrative style to neutral speech also has this effect in some cases. The mean value of 3,7 of the manipulated version of fragment 1 for example shows that quite some tension is experienced in fragment 1. The mean values of all fragments that contain climaxes show that participants experience tension in the manipulated fragments (the mean is above 3 for all fragments).

Comparing the mean answer values shows us that all manipulated fragments are considered to contain more tension than their neutral counterparts. For fragments 1 and 4, which do not contain climaxes, the difference is relatively high (difference of mean of 1,6 and 0,9). In case of fragments 6, 7 and 8 which do contain climaxes this difference is also high (1,3, 0,7 and 1,6). There are no remarkable values observable in the standard deviation and range values of the fragments.

Looking at the results of the t-test and Mann-Whitney test, neutral and manipulated versions of fragment 1, 6 and 8 turn out to be significantly different, with a significance that approaches zero. Fragment 4 and 7 both show a difference of means that is near acceptable significance of 0,05.

This section is concluded with a statistical analysis of the two participant groups. We will follow the same procedure as was used in the participant group comparison of the conversion rules evaluation, so we will compare the answers that are given for the three judgement criteria and the two kinds of fragments per group. The following table (table 10.5) shows the mean judgements of each participant group separated by fragment group and question.

participant group	1						2					
fragment group	a-fragments			b-fragments			a-fragments			b-fragments		
question	1	2	3	1	2	3	1	2	3	1	2	3
mean	3,1	2,6	2,3	3,3	3,0	3,2	2,9	3,0	2,4	3,6	3,3	3,3

Table 10.5. Group separated mean judgements

It is visible in these mean values that participant group 2 in general value the fragments a bit higher than participant group 1. The next step is to calculate with the t-test whether the judgement differences between groups are significant. So we compared the differences in means for all neutral fragments of group 1 with all neutral fragments of group 2, we did the same for the manipulated fragments (table 10.6).

fragment group	a-fragments			b-fragments		
question	1	2	3	1	2	3
significance	0,58	0,05	0,67	0,08	0,22	0,66

Table 10.6. Group separated difference significance levels.

We can see that in the case of judgement of naturalness of neutral fragments (a-fragments, question 2) there exists a significant difference in group judgement. There is almost significant difference in the judgement of the quality of storytelling of manipulated fragments (b-fragments, question 1). In both cases it is still the question whether this is a matter of group bias or that the difference is caused by the nature of the fragments.

For example, if we look back at the mean judgement values of the quality of storytelling of manipulated fragments is striking that there is one fragment that is rated remarkably low with respect to comparable fragments (table 10.3b, fragment 8b). So we may assume that the quality of storytelling of this fragment is low. If we take a look a the distribution of fragments among the two groups (§10.2), it turns out that this fragment was judged by participant group 1, which means the low judgement of this particular fragment influenced the average judgement value of manipulated fragments by group 1, consequently contributing to the almost significant group difference for the judgement of storytelling of this fragment group. So we can not say that there is a difference in group judgement here, because most likely this difference is caused by the fragments quality and not by a group bias.

10.5 Conclusion

Regarding the first judgement criterion we can say that in general there is an increase in the quality of storytelling after the conversion rules have been applied. This assertion is based on the observation that the absolute judgement means of most manipulated fragments are above average (on the 5-point Likert scale) with a considerable distance to the absolute judgement means of the neutral fragments, which are near the average. However, the degree of increase in the judgements is not large enough to prove mean difference with sufficient statistical ground, so the assertion is solely based on the observation of mean values.

In the consideration of naturalness only three out of eight manipulated fragments show a substantial increase in mean judgement of naturalness compared to the neutral version, of which one increase is significant. The other fragments of both neutral and manipulated nature have about equal naturalness judgements. Based on this there is no ground to say that naturalness is increased, so we conclude that the naturalness in general is equal for neutral and manipulated fragments.

Concerning the last judgement criterion there can be said that tension experience is not only increased by the presence of climax, but in some cases also the presence of narrative style increases the tension experience to a greater extent. All mean judgement values of the manipulated fragments are higher than that of neutral fragments, especially those of the climaxes. Three out of eight fragments have significantly different judgements values, of which two are climax fragments. The third climax fragment has judgements value differences that are near acceptable significance. In general we can say that the application of climaxes contributes to the experience of tension of speech. In a smaller extent this also applies for the narrative style.

The comparison of the two participant groups shows that in two fragment-question combinations a (near) significant difference occurs. The first is the case for the naturalness of neutral fragments, the second for the quality of storytelling of manipulated fragments. The first case, in which a significant difference occurs, is not really a cause for concern since the neutral fragments are only used to compare the answer values of the manipulated fragments to, the absolute judgements of the neutral fragments themselves is not the most interesting. On the other hand however, the almost significant difference in storytelling judgement of manipulated fragments is more worrying. If such a difference exists it is not possible to say whether this is caused by an actual group bias, or whether this is caused by the quality of the fragments. In this case however the cause of the difference could be traced back to the judgement of one fragment, which is judged considerably lower which could be caused by its quality. Therefore this difference can be contributed to this fragment's judgements and not to a group bias.

11 Conclusions

In this project we have developed a set of rules that can be applied to the prosodic representation of a neutrally spoken speech signal, resulting in narrative speech.

The first step in the process was to find out how prosodic functions are used by a storyteller to communicate a story. Most important prosodic functions used by a storyteller are the narrative speaking style and the realisation of climaxes. To find out how the realisation of those functions influence the speech signal, we performed an speech analysis in which we compared neutral speech and speech spoken by a storyteller. The analysis revealed that the acoustic features that are essential for the realisation of the prosodic functions are pitch, intensity and temporal aspects of the speech. Based on the observed behaviour of those features in narrative speech spoken by storytellers a set of rules was formulated which can be used to transform a neutrally spoken speech fragment in a narrative speech fragment. Evaluation of those rules revealed that the quality of storytelling is increased by the application of those rules, but not in such a high degree that this could be significantly proven. One important factor of negative influence on the quality and naturalness of storytelling is the fact that applying relatively large pitch and duration manipulations introduce distortion, caused by the signal processing method (both PSOLA and MBROLA).

Based on the conversion rules a module was constructed that automatically applies the conversion rules to generate narrative speech. The input of this module is an XML-file with the text that has to be synthesised, using tags to mark those positions that are of importance for the application of the prosodic functions. The module uses Fluency to determine the prosodic properties of a neutrally spoken version of the input text. Subsequently the conversion rules can be used to change these prosodic properties, resulting in manipulated prosodic information. This information is then used to have Fluency create the actual output of the module, the narrative pronunciation of the input text.

The implementation evaluation is the most important of all evaluations that were carried out because it evaluates the implementation of the results of all preceding phases. With respect to the other evaluations this evaluation is considered more important because it is carried out by a relatively large number of participants. The results of this evaluation show that apart from small exceptions there is a higher judgement of the manipulated fragments with regard to storytelling quality and perception of tension. In few cases this higher judgement is proven with acceptable significance, but in most cases the differences in judgement between the neutral and the manipulated versions of a fragment were not large enough to prove difference with sufficient significance.

We will now return to the goal of our project, which was to generate natural narrative speech. With regard to the *naturalness* of the speech that is generated we have seen in the conclusions of the implementation evaluation that no increase of naturalness is observed after the conversion rules have applied. On the other hand, the naturalness of manipulated fragments is not judged lower than that of the original, so the degree of naturalness that is achieved by Fluency is maintained.

The other aspect of the project goal was to create *narrative* speech. Based on the outcomes of the implementation evaluation we may conclude that the application of the conversion rules contributes to the narrative character of the speech (which we measured by examining storytelling quality and tension experience). Participants judge manipulated fragments to be of better storytelling quality and if a prosodic climax is present this increases the amount of tension experienced by the participant.

12 Recommendations

- For the purpose of creating more realistic narrative speech it is desirable that not only the prosodic functions that we focussed on in our project are studied, but also the paralinguistic functions used by a storyteller. Paralinguistic variation realises diversity in 'character voices', which means that different characters in a story can be represented by using a unique voice if they appear in the story. If more variation in paralinguistic features voice quality and voice qualification is realised more complete narrative speech is obtained. The addition of emotion to narrative speech will probably also increase the quality of storytelling.

- Based on the analysis phase results and the conversion rule evaluation we have determined the best values for some constants that are used in the conversion rules for narrative speech. The use of constant values here doesn't always yield enough variation in speech; the use of a more probabilistic approach may result in a higher degree of naturalness.

- The placement of sentence accents falls under the responsibility of the annotator, or in the in context of the Virtual Storyteller project under the responsibility of the plot generator. For the determination of the position of those sentence accents that should be increased in duration a more elaborate study of the underlying grammatical model is needed. This determination can best be done during the language generation (which creates a natural language narrative based on an abstract representation of the plot) because at that point the grammatical structure of the sentences is known.

13 References

[1] Rietveld, A.C.M. en Heuven, V.J. van, (1997), “Algemene fonetiek”, Coutinho

[2] http://www.rhetorica.net/textbook/, “Rhetorica, a rhetoric primer”, A. R. Cline, Park University, Parkville, USA

[3] Cowie, R., “Describing the Emotional States Expressed in Speech”, *ISCA Workshop on Speech & Emotion, Northern Ireland 2000*, p. 11-18.

[4] Cornelius, R. R., “Theoretical Approaches to Emotion”, *Proc. of ISCA Workshop on Speech and Emotion,* Belfast, September 2000.

[5] Roach, P. (2000). “Techniques for the phonetic description of emotional speech”. Proceedings of the ISCA Workshop on Speech and Emotion. Newcastle, Northern Ireland. September 2000. 53-59.

[6] Cahn, J. “Generation of Affect in Synthesized Speech”, In Proceedings of AVIOS 89, pp. 251-256, 1989.

[7] Schroder, M. “Emotional Speech Synthesis--a Review”. *Proceedings of Eurospeech 2001.* Aalborg. pp.561-564.

[8] Theune, M., Faas, S., Nijholt, A., Heylen, D., “The Virtual Storyteller”, *ACM SIGGROUP Bulletin*, Volume 23, Issue 2, ACM Press, pages 20-21.

[9] Fackrell, Vereecken, Buhmann, Martens, Van Coile, “Prosodic variation with text type”, Proceedings 6th conference ICSLP 2000. Vol. 3. 2000. pp. 231-234

[10] Murray, I.R., Arnott, J.L., “Implementing and testing of a system for producing emotion-by-rule in synthetic speech”, Speech Communication, 16, 1995, pp. 369-390, Dundee

[11] Trochim, W.M., “Research Methods Knowledge Base”, http://trochim.human.cornell.edu/kb/, Cornell University

[12] Jurafsky, D., Martin, J.H., “Speech and language processing, an introduction to natural language processing, computational linguistics and speech recognition”, Chapter 7.8, Prentice Hall, 2001

[13] Ruttkay Zs., van Moppes, V., Noot, H., “The jovial, the reserved and the robot”, AAMAS2003 Conference

[14] Mixdorff H.(2002), “Speech Technology, ToBI, and Making Sense of Prosody”, in Proc. SpeechProsody2002, Aix-en-Provence, 31-37.

[15] Mozziconacci, S. “Prosody and emotions”, in Proc. SpeechProsody2002, Aix-en-Provence

[16] Murtaza Bulut, Shrikanth Narayanan and Ann Syrdal, ``Expressive speech synthesis using a concatenative synthesizer", Proc. of ICSLP, (Denver, CO), 2002

[18] Pierrehumbert, J., Hirschberg, J. (1990). “The Meaning of Intonation in the Interpretation of Discourse.” In P. Cohen, J. Morgan, and M. Pollack, (eds.) *Intentions in Communication.* MIT Press, Cambridge MA. 271-311.

[19] Link, K. E., Kreuz, R. J., Graesser, A. C., and the Tutoring Research Group (2001). “Factors that influence the perception of feedback delivered by a pedagogical agent.” *International Journal of Speech Technology, 4*, 145-153.

[20] http://wwwbox.uni-mb.si/eSpeech/, Emotional speech Group at DSPLAB, University of Maribor, Faculty of Electrical Engineering and Computer Science, Slovenia.

[21] J. Cassell, C. Pelachaud, N. Badler, M. Steedman, , T. Becket, B. Douville, S. Prevost, M. Stone (1994). “ANIMATED CONVERSATION: Rule-based Generation of Facial Expression, Gesture & Spoken Intonation for Multiple Conversational Agents.” ACM-SIGGRAPH, 413-420. *Cybernetics* **23**: 3, pp. 665-685.

[22] http://www.dfki.de/~baldes/rocco/personality/, “Literature survey about personality and emotions”, Baldes S., DFKI GmbH, Germany

[23] Zetterholm, E. 1998. “Prosody and voice quality in the expression of emotions”. *Proceedings of the Seventh Australian International Conference on Speech Science and Technology*, 109-113. Sydney, Australia.

[24] http://www.cstr.ed.ac.uk/projects/sable/, “The Sable Consortium”, The Centre for Speech Technology Research, University of Edinburgh, Scotland

[25] http://www.w3.org/, World Wide Web Consortium

[26] http://www.w3.org/TR/2003/CR-speech-synthesis-20031218/, “Speech Synthesis Markup Language Version 1.0”, World Wide Web Consortium

[27] Kshirsagar, S., Guye-Vuilleme, A., Kamyab, K., “Avatar Markup Language”, Proceedings of the 8th Eurographics Workshop on Virtual Environments, pp 169-177., May, 2002

[28] http://www.w3.org/DOM/, “W3C Document Object Model”, World Wide Web Consortium

[29] Bozkurt, B., Dutoit, T., “An implementation and evaluation of two diphone based synthesizers for Turkish”, Proceedings of 4th ISCA Tutorial and Research Workshop on Speech Synthesis, pp.247-250, Blair Atholl, Scotland, 2001

[30] Theune, M., “From data to Speech, language generation in context”, PhD thesis, Eindhoven University of Technology, 2000

[31] http://nextens.uvt.nl/, “Nextens: open source Text-to-Speech for Dutch”, Department of Language and Speech, University of Nijmegen, Induction of Linguistic Knowledge Group, University of Tilburg, The Netherlands

[32] http://www.fluency.nl, “Fluency Home Page”, Fluency, Van Dale Lexicografie, The Netherlands

[33] http://www.praat.org/, “Praat: doing phonetics by computer”, Boersma, P., Weenink, D., Institute of Phonetic Sciences, University of Amsterdam, The Netherlands

[34] http://www.koenmeijs.nl/code/, program code used for implementing a module that creates narrative speech based on a certain input text.

14 Appendices

14.1 Appendix A: Fragments used in analysis

A.1 Fragments used for analysis of pitch and intensity of narrative style

Fragment: ***News_1, Dutch Radio 1 News 26 September 2003, 17.00u, spoken by Onno Duyvené de Wit, Dutch male newsreader***
“Veel klanten twijfelen aan veiligheid internet bankieren, ook onderzoeken Nederland naar netwerk kinderporno en britse zanger Robert Palmer overleden. Rekeninghouders vinden internetbankieren nog steeds niet veilig. Ruim een vijfde van hen twijfelt aan de betrouwbaarheid van de computertechniek.”

Fragment: ***News_2, Dutch Radio 1 News, 21 September 2003, 18.00u, spoken by Onno Duyvené de Wit, Dutch male newsreader***
“Zo’n vijftig gemeenten hebben meegedaan aan de autoloze zondag.”

Fragment: ***News_3, Dutch Radio 1 News, 21 September 2003, 18.00u, spoken by Onno Duyvené de Wit, Dutch male newsreader***
"De officiële start was in de Huygenslaan in Arnhem, een drukke verkeersader midden in een wijk waar veel kinderen spelen."

Fragment: ***Child_1, “Klaas Vaak”, “Luister sprookjes en vertellingen”, Lekturama.***
"...liepen door een lange gang, die zo laag was dat Jelmar ervoor moest oppassen zijn hoofd niet te stoten"

Fragment: ***Child_2, “Klaas Vaak”, “Luister sprookjes en vertellingen”, Lekturama.***
“De muizen, maar ook Jelmar lieten zich het eten goed smaken. Alleen kon Jelmar niet veel op.”

Fragment: ***Child_3, “Klaas Vaak”, “Luister sprookjes en vertellingen”, Lekturama.***
“Maar omdat ie klein getoverd was had ie natuurlijk ook maar een heel klein maagje.”

Fragment: ***Adult_1, ”De eekhoorn en de mier”, Toon Tellegen, Dutch male storyteller***
“Toen de mier weer eens een verre reis maakte, zat de eekhoorn voor zijn raam, en dacht aan hem.”

Fragment: ***Adult_2, ”De eekhoorn en de mier”, Toon Tellegen, Dutch male storyteller***
“Plotseling begon hij te rillen en dacht: ‘bestaat de mier eigenlijk wel?’ Hij ging aan zijn tafel zitten, en verborg zijn hoofd in zijn handen.”

A.2 Fragments used for analysis of pause length of narrative style

newsreader

'.. nog steeds niet veilig __ en ruim een vijfde van hen ..'
pause duration 0,283 sec
'.. was in de Huygenslaan in Arnhem, __ een drukke verkeersader ..'
pause duration 0,401 sec
'Zijn hoofdpersonen zijn vaak eenzame mensen, __ die toch worden geraakt ..'
pause duration 0,313 sec
'..de democratische alliantie zegt, __ dat alle Zuid-Afrikanen trots zijn.'
pause duration 0,299 sec

'.. Zuid-Afrikanen trots zijn. __ Het werk van de drieënzestig-jarige Coetzee ..'
pause duration 0,431 sec
'.. door de apartheidspolitiek. __ Zijn hoofdpersonen ..'
pause duration 0,476 sec
'.. door de wereld om hen heen. __ In Nederland is veel belangstelling..'
pause duration 0,692 sec
'.. 65.000 exemplaren verkocht. __ Zijn nieuwste werk..'
pause duration 0,508 sec

child storyteller

'.. de muizen, __ maar ook Jelmar ..'
pause duration 0,384 sec
'.. oogleden worden dan zwaar, __ en je beleeft de wonderlijkste dingen.'
pause duration 0,300 sec
'Piep-piep en Jelmar liepen door een lange gang, __ die zo laag was dat ..'
pause duration 0,549 sec
'.. weer naar bed, __ nadat Klaas Vaak hem weer ..'
pause duration 0,569 sec

'.. komt Klaas Vaak. __ Hij is het grappigste mannetje ..'
pause duration 1,32 sec
'.. dat je je kunt voorstellen. __ Hij draagt een zak met toverzand.'
pause duration 1,28 sec
'.. toverzand. __ Onder het vertellen ..'
pause duration 1,60 sec
'..zonder dat je het merkt. __ Je oogleden worden ..'
pause duration 1,01 sec

A.3 Fragments used for analysis of vowel duration of narrative style

e:_newsreader:
'zo'n vijftig gemeenten hebben m_EE_gedaan aan de autoloze zondag'
vowel duration 0,137 sec
'Britse zanger Robert Palmer overl_EE_den'
vowel duration 0,153 sec
'Tw_EE_derde van de thuisbankiers neemt zelf maatregelen'
vowel duration 0,117 sec
e:_childstoryteller:
'natuurlijk ook maar een h_EE_l klein maagje'
vowel duration 0,110 sec
'All_EE_n kon Jelmar niet veel op'
vowel duration 0,143 sec
'.. en je bel_EE_ft de wonderlijkste dingen'
vowel duration 0,128 sec
'hij voelde zich soms zo all_EE_n in zijn mooie huis'
vowel duration 0,200 sec

o:_ newsreader:
'volgens de organisatie een sch_OO_lvoorbeeld van de problemen die veel mensen tegenkomen'
vowel duration 0,123 sec
'KPN vers_O_bert salaris topman Scheepbouwer'
vowel duration 0,128 sec
'.. zo levert topman Scheepbouwer een vaste b_O_nus ..'
vowel duration 0,162 sec
o:_ childstoryteller:
'v_OO_r de kinderen gaan slapen'
vowel duration 0,186 sec
'liepen door een lange gang, die z_O_ laag was'
vowel duration 0,174 sec
'hij voelde zich soms z_O alleen in zijn mooie huis'
vowel duration 0,186 sec

O_ newsreader:
'hebben meegedaan aan de autoloze z_O_ndag'
vowel duration 0,081 sec
'gew_O_nde bij metro-ongeluk in Londen'
vowel duration 0,056 sec
'KPN verlaagt het salaris van z'n t_O_pbestuurders'
vowel duration 0,061 sec
O_ childstoryteller:
'en je beleeft de w_O_nderlijkste dingen'
vowel duration 0,147 sec
'De krokussen staken hun k_O_pjes boven de grond..'
vowel duration 0,054 sec
'De krokussen staken hun kopjes boven de gr_O_nd..'
vowel duration 0,136 sec

a:_ newsreader:
'er zijn een p_AA_r zaken die je als klant in de gaten moet houden'
vowel duration 0,162 sec
'KPN verl_AA_gt het salaris van zijn topbestuurders'
vowel duration 0,100 sec
'm_AA_r volgens KPN ..'
vowel duration 0,094 sec
a:_ childstoryteller:
'komt Kl_AA_s v_AA_k'
vowel duration 0,114 sec

vowel duration 0,129 sec
‘Ze trokken de mooiste kleren _AA_n ..’
vowel duration 0,184 sec
‘.. en liepen trots door de schitterende k_A_mers.’
vowel duration 0,175 sec

i._ newsreader:
'... n_IE_uwe leider van de eredivisie'
vowel duration 0,089 sec
‘Bovend_IE_n krijgen topbestuurders die bij KPN vertrekken...’
vowel duration 0,080 sec
‘Nelie Kroes, d_IE sinds kort voorzitter is ...’
vowel duration 0,092 sec
i._ childstoryteller:
'Piep-p_IE_p en Jelmar liepen door een lange gang'
vowel duration 0,086 sec
‘.. het maakte hem eigenlijk wat verdr_IE_tig’
vowel duration 0,095 sec
‘.. de andere feeën gaven de prinses een mooie stem, vr_IE_ndelijkheid, gezondheid,..
vowel duration 0,112 sec

A_ newsreader:
'de officiele start was in de huygenslaan in _A_rnhem'
vowel duration 0,104 sec
‘een vaste bonus van een h_A_lf miljoen euro in’
vowel duration 0,069 sec
‘verder worden financiële voordelen geschr_A_pt’
vowel duration 0,087 sec
A_ childstoryteller:
'hij was het gr_A_ppigste mannetje dat je je kan voorstellen'
vowel duration 0,066 sec
‘maar in de tuin van de reus w_A_s het nog winter’
vowel duration 0,087 sec
‘de sneeuw was uit zijn tuin verdwenen en A_lle bomen stonden in bloei.’
vowel duration 0,155 sec

E_ newsreader:
'.. neemt z_E_lf maatregelen om de internettransacties te beveiligen'
vowel duration 0,095 sec
‘.. konden oplopen tot zo’n E_lf miljoen’
vowel duration 0,067 sec
‘dat hem r_E_cht gaf op een extra jaarinkomen ..’
vowel duration 0,071 sec
E_ childstoryteller:
'.. dat J_E_lmar ervoor moest oppassen zijn hoofd niet te stoten'
vowel duration 0,063 sec
‘in deze tuin wordt het nooit l_E_nte’
vowel duration 0,100 sec
‘hij sprong uit b_E_d en keek naar buiten’
vowel duration 0,094 sec

A.4 Absolute and relative vowel durations

Newsreader

syllable	Zo'n	vijf	tig	ge	meen	ten	heb	ben	mee	ge	daan
vowel	o:	Ei	I	@	e:	@	E	@	e:	@	a:
vowel kind	long	long	short	short	long	short	short	short	long	short	long
absolute duration	0,06	0,12	0,05	0,03	0,13	0,08	0,09	0,08	0,09	0,03	0,10
relative duration	0,59	1,19	0,84	0,47	1,35	1,30	1,55	1,31	0,90	0,52	0,97

Child storyteller

syllable	lie	pen	door	een	lan	ge	gang	die	zo	laag	was	dat	jel	mar
vowel	i.	@	o:	@	A	@	A	i.	o:	a:	A	A	E	A
vowel kind	long	short	long	short	short	short	short	long	long	long	short	short	short	short
absolute duration	0,10	0,03	0,07	0,09	0,08	0,06	0,08	0,08	0,17	0,22	0,11	0,08	0,07	0,13
relative duration	0,80	0,40	0,54	0,74	0,92	0,74	1,02	0,69	1,40	1,83	1,35	1,01	0,83	1,63

14.2 Appendix B: Constant Evaluation Fragments

Below you find an enumeration of the fragments that were created for the constant evaluation. First the positions of the sentence accents are provided, including the start and end time of the syllable. Then for each version of the fragment (*a* and *b*) the distinguishing characteristics are given. Fragment 1 until and including 11 are based on the narrative speaking style; the rest is based on the climaxes. All climax sentences were first manipulated according to the narrative style rules; afterwards the climax rules were applied.
The general procedure that was followed is that the constants from table 7.2 in paragraph 7.3 are handled one by one and used as a basis for a fragment. For every new question the first fragment is used as a baseline to compare the second to. The second then has a new value for one of the evaluation constants. For every fragment only the changes with respect to its predecessor are given, so its remaining characteristics are the same as its predecessor.

Question 1

```
"De boom moest worden omgehakt met een grote bijl"
accent     _                _                  _    _
start      0,32             1,32               2,47 2,99
end        0,54             1,53               2,72 3,34
```

Frag_1_a
Original Fluency pronunciation
Frag_1_b
Pitch: $m_1 = 0$
$m_2 = 0,75$
dmpi=40

Question 2

```
            "Jelmar liep door een lange gang, die zo laag was dat hij
accent        -                     -              -   -
start         0,01                  1,23           2,55 2,80
end           0,34                  1,46           2,80 3,16

bijna zijn hoofd stootte."
accent      -           -
start       3,87        4,61
end         4,01        4,90
```

Frag_2_a
Original Fluency pronunciation
Frag_2_b
Pitch: $m_1 = 0$
$m_2 = 0,75$

dmpi=60

Question 3

Frag_3_a = Frag_2_b
Frag_3_b
Pitch: m_1 = 0,25
m_2 = 0,50
dmpi=60

Question 4

```
      "Hij was de rijkste man van het hele land en toch was hij niet
accent            -                   -            -            -
start             0.59                1.77         2.78         3.39
end               0.87                2.06         2.94         3.71

blij en gelukkig."
accent          -
start           4.28
end             4.39
```

Frag_4_a
Original Fluency pronunciation
Frag_4_b
Pitch: m_1 = 0
m_2 = 0,75
dmpi=40

Question 5

Frag_5_a = Frag_4_b
Frag_5_b

Intensity: c = 4dB
k = 0,2 sec

Question 6

Frag_6_a = Frag_4_b

Frag_6_b
Intensity: c = 6dB
k = 0,2 sec

Question 7

```
      "De dochters hadden het kloppen gehoord, en achter de deur
accent     -                    -                 -
start      0.27                 1.15              2.09
end        0.34                 1.33              2.23
```

```
stonden zij stilletjes te luisteren"
accent              -            -
start               3.45         4.09
end                 3.66         4.35
```

Frag 7_a
Original Fluency pronunciation

Frag 7_b
Pitch: m_1 = 0
m_2 = 0,75
dmpi=30

Question 8

Frag 8_a = Frag 7_b
Frag 8_b
Intensity: c = 2dB
k = 0.2 sec

Question 9

Frag 9_a = Frag 7_b
Frag 9_b
Intensity: c = 2dB
k = 0 sec

Question 10

Frag 10_a = Frag 7_b
Frag 10_b
Intensity: c = 4dB
k = 0 sec

Question 11

Frag 11_a = Frag 10_b
Frag 11_b
lengtened, factor 1.2
inserted pause

Question 12

“Iedereen wachtte in stilte *en toen* klonk er een daverende knal”

Frag 12_a
pitch shift “toen”: 80Hz, afterwards: constant
duration “toen” * 2,56

Frag 12_b
pitch shift “toen”: 120Hz, afterwards: constant
duration * 2,56

Question 13

Frag 13_a = Frag 12_a
Frag 13_b
pitch shift “toen”: 80Hz, afterwards: rising 100Hz
duration “toen” * 2,56

Question 14

Frag 14_a = Frag 12_a
Frag 14_b
pitch shift “toen”: 80Hz, afterwards: constant
duration “toen” * 2,56
intensity: rise 6dB, decreasing

Question 15

Frag 15_a = Frag 12_a
Frag 15_b
pitch shift “toen”: 80Hz, afterwards: constant
duration “toen” * 2,56
intensity: rise 10dB, decreasing

Question 16

Frag 16_a = Frag 12_a
Frag 16_b
pitch shift “toen”: 80Hz, afterwards: constant
duration “toen” * 2,56
intensity: rise 10dB, staying

Question 17

Note: This fragment consists of two sentences, the first is to build up the tension, the second contains the climax.

```
“Tree voor tree beklom de prins de trap tot hij uiteindelijk een”
accent   -         -             -                       -
start    0.07      0.66          1.59                    2.99
end      0.34      0.95          1.85                    3.19
```

```
“grote houten deur bereikte.”
accent   -
start    3.66
end      3.88
```

```
“Hij deed de deur open en... daar lag de slapende prinses.”
accent      -       -   -    -      -            -        -
start_syl   0.32  0.75  1.07 1.48   2.65         3.37     4.04
end_syl     0.49  1.02  1.24 1.64   2.90         3.59     4.23
d_pit_inc   50    55    60   80     50           25       25
n           2.2   2.0   1.83 1.375  2.2          4.4      4.4

start_vow   0.38  0.83  1.07 1.50   2.74
end_vow     0.49  0.98  1.24 1.57   2.87
fact_dur    1.3   1.3   1.5  2.0    1.7
```

Frag 17_first

Pitch: $m_1 = 0$

$m_2 = 0,75$

dmpi=40

Frag 17_second_a

see above, pitch and duration of vowels are gradually increased

Frag 17_second_b

```
“Hij deed de deur open en... daar lag de slapende prinses.”
accent      -       -   -    -      -            -        -
d_pit_inc   25      30    30   60   30           25       25
n           4.4     3.66  3.66 1.83  3.66        4.4      4.4
```

Question 18

Frag 18_a = Frag 17_second_b

Frag 18_b

Duration in climax:

```
“Hij deed de deur open en... daar lag de slapende prinses.”
accent      -       -   -    -      -            -        -
fact_dur    1.1     1.2 1.3  1.5    1.7
```

Question 19

Frag_19_a

Original Fluency pronunciation

Frag_19_b

```
“Voetje voor voetje sprong hij over de stenen heen, gleed weg”
accent  -           -                          -            -
start   0.07        0.63                       2.11         3.15
end     0.21        0.78                       2.34         3.41
dmpi    40          40                         40           40
n       2.75        2.75                       2.75         2.75
```

```
“over de laatste steen en.. plonste in het water.”
accent -     -  -        -   -     -             -
start  3.79  4.16 4.22   4.86 5.13  5.35          6.09
end    3.96  4.22 4.45   5.13 5.26  5.57          6.32
dmpi   40    40   50     60   70    50            40
n      2.75  2.75 2.2    1.83 1.57  2.2           2.75
```

Add pause after “en”

Question 20

Frag_20_a = Frag_19_b

Frag_20_b

Duration manipulation:

```
“over de laatste steen en.. plonste in het water.”
accent -     -  -        -   -     -
s_vow  3.80     4.29     4.90 5.12  6.22
e_vow  3.97     4.46     5.06 5.21  6.32
f_dur  1.3      1.5      1.7  2.0   2
```

Question 21

Frag_21_a = Frag_19_a

Frag_21_b

```
“over de laatste steen en.. plonste in het water.”
accent -     -  -        -   -     -             -
dmpi   40    50   60     80   90    60            40
n      2.75  2.2  1.83   1.38 1.22  1.83          2.75
```

Question 22

Frag_22_a = Frag_20_b

Frag_22_b

```
“over de laatste steen en.. plonste in het water.”
accent -     -  -        -   -      -             -
s_vow  3.80     4.29     4.90 5.12  5.42
e_vow  3.97     4.46     5.06 5.21  5.52
f_dur  1.3      1.5      1.7  2.0   1.5
```

Question 23

```
            "Jelmar liep door een lange gang, die zo laag was dat hij
accent        -                     -              -   -
start         0,01                  1,23           2,55 2,80
end           0,34                  1,46           2,80 3,16

bijna zijn hoofd stootte."
accent      -           -
start       3,87        4,61
end         4,01        4,90
```

Frag_23_a

Pitch: m_1 = 0

m_2 = 0,75

dmpi = 30

Frag_23_b

Duration increase on accented vowels of ‘zo’ and ‘laag’

Duration factor = 1.5

14.3 Appendix C: Statistics of constant evaluation

The following table lists the average and standard deviation of the answers that the participants gave for each question. The possible answers a participant could give for a question were "fragment 1 is best", "equal" and "fragment 2 is best", corresponding with values 0, 1 and 2. The last two rows show the average number of times that fragment 1 and 2 were listened to.

question	1	2	3	4	5	6	7	8	9	10	11	12
average	1,0	1,2	0,8	1,6	0,4	0,4	0,8	1,0	0,8	1,2	0,4	1,0
st. dev	1,1	1,0	0,4	0,9	0,5	0,5	1,1	0,7	0,5	0,8	0,9	1,2
#times frag 1	1,9	1,7	1,2	1,4	1,8	1,8	1,6	2,2	1,2	1,4	1,2	2,4
#times frag 2	1,4	1,7	1,2	1,8	1,8	1,4	1,8	1,8	1,4	1,4	1,2	2

question	13	14	15	16	17	18	19	20	21	22	23
average	1,0	1,4	1,4	0,6	1,3	0,8	1,8	1,0	1,8	1,2	1,7
st. dev	1,2	0,5	0,5	0,9	1,0	0,4	0,5	1,2	0,5	0,4	0,9
#times frag 1	1,6	1,4	1,4	1,6	1,8	2,4	1,6	1,6	1,0	2,0	1,5
#times frag 2	1,4	1,4	1,4	1,8	1,6	1,6	1,8	1,6	1,2	1,6	1,5

14.4 Appendix D: Conversion Rule Evaluation Fragments

This section lists the accent positions, start time of syllables and end time of syllables of all fragments that are used in the evaluation.

frag_1_a

```
             Als ze moe waren namen ze een heerlijk warm bad in één van de
accent               -                       -       -    -      -
start                0.31                    1.60    2.17 2.56   2.94
end                  0.57                    1.91    2.50 2.65   3.15
lvow_st                                      1.70                 2.94
lvow_end                                     1.86                 3.09
             gouden badkuipen.
accent               -
start                3.94
end                  4.10
```

frag_2_a

```
             Zes weken lang moest hij gaan reizen want er waren hele
accent        -         -                                        -
start         0.03      0.75                                     3.02
end           0.32      1.04                                     3.31
lvow_st                                                          3.10
lvow_end                                                         3.30

             belangrijke zaken te doen.
accent                     -
start                      4.22
end                        4.49
lvow_st
lvow_end
```

frag_3_a

```
       Op een dag reed hij naar zijn buurvrouw die twee lieve mooie dochters had
accent -                                                -            -
start   0.01                                            2.79         3.55
end     0.09                                            2.98         3.72
```

frag_4_a

```
       Maar hij kon niet veel op want hij had maar een heel klein maagje.
accent               -        -                         -     -
start               0.71      1.27                      2.82  3.13
end                 0.97      1.42                      3.05  3.47
lvow_st                                                 2.90  3.25
lvow_end                                                3.02  3.41
```

frag_5_a

```
       Hij had wel een hele grote mond voor zo’n klein ventje.
accent                 -           -         -
start                  0.94        1.83      2.45
```

```
end                    1.21        2.26      2.69
lvow_st                1.01                  2.52
lvow_end               1.21                  2.65
```

climax_1
frag_6_a

pitch rise on climax word is performed only once (so no narrative style pitch adaption)

```
       De rust leek teruggekeerd maar toen klonk er een daverende knal.
accent     -                            C               -          -
start      0.22                         2.00            2.97       3.70
end        0.40                         2.17            3.20       4.04
lvow_st                                                 3.08
lvow_end                                                3.20
```

frag_7_a

```
       De kust leek veilig tot er plotseling iemand voor hem opdook.
accent     -                       C         -
start      0.28                    1.80      2.41
end        0.45                    2.06      2.53
```

frag_8_a

used notation: CB= Climax Begin, CT = Climax Top, CE = Climax End

```
       Het bos leek bij iedere stap donkerder te worden. Maar in de verte zagen
accent      -           -                                 -          -
start       0.15        0.87                              3.55       4.09
end         0.40        0.96                              3.79       4.33
vow_st                  0.87
vow_end                 0.96
```

```
       ze een vaag lichtschijnsel. Dichter en dichterbij kwamen ze, nog maar
accent         -                   -          -                          CB
start          5.26                7.34       7.98                       9.83
end            5.56                7.60       8.17                       10.08
pitch                                                                    +10
vow_st         5.32                                                      9.91
vow_end        5.48                                                      10.00
dur                                                                      1.1
```

```
       enkele stappen te gaan en... Daar zagen ze een prachtige waterval in
accent                       CT      -                 -
start  10.08  10.52 11.07 11.18 11.50 12.14             13.22
end    10.20  10.78 11.17 11.50 11.80 12.43             13.44
pitch  +20    +30   +40   +50   +60  +25                +15
vow_st 10.08  10.69 11.10 11.28 11.51 12.25
vow_en 10.14  10.78 11.18 11.42 11.64 12.42
dur    1.1    1.2   1.3   1.4   1.5   1.5
```

```
       volle zon.
accent  -    CE
```

```
start   14.57
end     14.77
pitch    +15
```

14.5 Appendix E: Implementation Evaluation Fragments

The following fragments were used in the implementation evaluation. First each fragment is given in plain text, followed by the annotated version of the fragments.

fragment_1 Er was eens een man die geweldig rijk was.
fragment_2 Dan zat hij in een grote stoel met een schitterend geborduurde rug.
fragment_3 Hij was de rijkste man van het hele land en toch was hij niet blij en gelukkig.
fragment_4 Die baard maakte hem zo afschuwelijk lelijk dat iedereen op de loop ging zodra hij in de buurt kwam.
fragment_5 Als ze maar dachten dat ze ergens muziek hoorden dan bewogen ze zich sierlijk op de maat van die muziek.
fragment_6 Hij rende zo hard als hij kon maar toen struikelde hij over zijn eigen benen.
fragment_7 Hij wilde zich omkeren en toen klonk er plotseling een harde knal.
fragment_8 Stap voor stap kwam hij dichterbij. Toen hij haar dicht genoeg genaderd was greep hij haar bij haar keel en toen bleek ze plotseling verdwenen.

```
fragment_1   <s>Er <sentence_accent extend="no">was</sentence_accent> eens een man die
             ge<sentence_accent extend="yes">wel</sentence_accent>dig rijk was.</s>
fragment_2   <s>Dan zat hij in een <sentence_accent extend="no">gro</sentence_accent>te
             stoel met een <sentence_accent extend="no">schit</sentence_accent>terend
             geborduurde rug.</s>
fragment_3   <s>Hij was de <sentence_accent extend="no">rijk</sentence_accent>ste man
             van het <sentence_accent extend="yes">he</sentence_accent>le land en
             <sentence_accent extend="no">toch</sentence_accent> was hij niet blij en
             gelukkig.</s>
fragment_4   <s>Die baard maakte hem <sentence_accent extend="yes">zo</sentence_accent>
             afschuwelijk lelijk dat <sentence_accent
             extend="no">ie</sentence_accent>dereen op de loop ging zo<sentence_accent
             extend="no">dra</sentence_accent> hij in de buurt kwam.</s>
fragment_5   <s>Als ze maar <sentence_accent extend="no">dach</sentence_accent>ten dat
             ze ergens muziek hoorden dan bewogen ze zich <sentence_accent
             extend="yes">sier</sentence_accent>lijk op de <sentence_accent
             extend="no">maat</sentence_accent> van die muziek.</s>
fragment_6   <s>Hij rende zo hard als hij kon <climax type="imediate">maar toen</climax>
             struikelde hij over zijn <sentence_accent
             extend="yes">ei</sentence_accent>gen benen.</s>
fragment_7   <s>Hij wilde zich omkeren <climax type="imediate">en toen</climax> klonk er
             <sentence_accent extend="no">plotseling</sentence_accent> een harde
             knal</s>
fragment_8   <s><sentence_accent extend="no">Stap</sentence_accent> voor stap kwam hij
             <sentence_accent extend="no">dich</sentence_accent>terbij.</s>
             <s>Toen hij haar dicht genoeg genaderd was greep hij <climax
             type="increasing">haar bij haar keel en <climax_top/> toen bleek ze
             plotseling verdwenen.</climax></s>
```

14.6 Appendix F: Introduction text used in evaluation environment

The Dutch introduction text below was showed to the participant before an experiment took place. This introduction text was specifically used as introduction to the experiment of the implementation evaluation; the experiments of other evaluations had similar introduction text.

Welkom bij deze evaluatie van verhalende spraak. In deze evaluatie zul je meerdere spraakfragmenten te horen krijgen die beoordeeld moeten worden. De fragmenten zijn allemaal met de computer gecreëerd door middel van spraaksynthese. Na creatie zijn de fragmenten gemanipuleerd met het doel de spreker te laten klinken als een verhalenverteller.

De werkwijze van de evaluatie is als volgt:

Je krijgt telkens één geluidfragment aangeboden dat je dient te beluisteren. Bij elk fragment worden drie vragen gesteld, die ieder op een 5-punts schaal beantwoord dienen te worden.

Er zijn in totaal 8 fragmenten. Je mag een fragment meerdere keren beluisteren, maar niet vaker dan 3 keer.

Er wordt tevens de mogelijkheid geboden om in een tekstvlak extra opmerkingen bij de fragmenten te plaatsen, wanneer je iets opvalt dat het vermelden waard is.

In het belang van het onderzoek vragen wij niet te schromen kritisch te zijn in je beoordelingen. Denk niet dat je overal spanning of verhalende spraak zou moeten waarnemen want dat hoeft niet het geval te zijn.

Het is goed van te voren te beseffen dat door de computer gegenereerde spraak niet perfect klinkt en soms slecht verstaanbaar is. Het is dus niet de bedoeling dat je de fragmenten beoordeelt op verstaanbaarheid, maar op de kwaliteit van vertellen. Om een idee te geven van hoe synthetische spraak klinkt worden hieronder eerst een neutrale spraakfragmenten aangeboden dat beluisterd moet worden voordat je verder kunt gaan met de evaluatie.

Beluister eerst het volgende fragment om gewend te raken aan artificial speech:

"Ze aten en dronken en dansten op het groene grasveld. Als ze moe waren namen ze een heerlijk warm bad in één van de gouden badkuipen en zodra ze zich opgeknapt hadden gingen ze weer aan tafel. Nooit hadden ze zo'n plezier gehad."

Klik vervolgens op onderstaande button om verder te gaan naar de eerste vraag.