University of Twente
Enschede - The Netherlands

Virtual Storytelling: Emotions for the narrator

Master's thesis, August 2007

H.A. Buurman

Committee
dr. M. Theune
dr. ir. H.J.A. op den Akker
dr. R.J.F. Ordelman

Faculty of Human Media Interaction,
Department of Electrical Engineering,
Mathematics & Computer Science,
University of Twente

Preface

During my time here as a student in the Computer Science department, the field of language appealed to me more than all the other fields available. For my internship, I worked on an assignment involving speech recognition, so when a graduation project concerning speech generation became available I applied for it. This seemed like a nice complementation of my experience with speech recognition. Though the project took me in previously unknown fields of psychology, linguistics, and (unfortunately) statistics, I feel a lot more at home working with- and on Text-to-Speech applications.

Along the way, lots of people supported me, motivated me and helped me by participating in experiments. I would like to thank these people, starting with Mariët Theune, who kept providing me with constructive feedback and never stopped motivating me. Also I would like to thank the rest of my graduation committee: Rieks op den Akker and Roeland Ordelman, who despite their busy schedule found time now and then to provide alternative insights and support. My family also deserves my gratitude for their continued motivation and interest. Lastly I would like to thank all the people who helped me with the experiments. Without you, this would have been a lot more difficult.

Herbert Buurman

Samenvatting

De ontwikkeling van het virtueel vertellen van verhalen staat nooit stil. Zolang een virtuele verhalenverteller niet op een vergelijkbaar niveau presteert als een menselijke verhalenverteller is er nog wat aan te verbeteren. De Virtual Storyteller van de HMI afdeling van de Universiteit Twente gebruikt een Text-to-Speech programma dat ingevoerde tekst automatisch omzet naar spraak.

Het zeker niet triviaal om deze automatisch gegenereerde spraak menselijk te laten klinken, vooral als het om het vertellen van verhalen gaat. Er zijn zoveel aspecten van het vertellen waar vertellers hun stem gebruiken om de act te verbeteren. Dit verslag beschrijft het onderzoek waarbij gekeken wordt hoe verhalenvertellers hun stem gebruiken om emoties van personages in het verhaal over te brengen. Er wordt specifiek gekeken naar hoe de verhalenverteller zijn stem verandert om een personage in het verhaal iets te laten zeggen op een emotionele manier.

Een experiment is uitgevoerd om de emotionele lading van spraakfragmenten van personages te identificeren. Deze fragmenten worden daarna geanalyseerd in een poging om uit te vinden hoe de emotionele ladingen gekoppeld zijn aan de manier waarop de verhalenverteller zijn stem gebruikt. De analyse wordt daarna gebruikt bij het opstellen van een model dat het Text-to-Speech programma gebruikt om in plaats van neutrale spraak emotionele spraak te genereren.

Het model is geimplementeerd in een open-source Text-to-Speech programma dat een Nederlandse stem gebruikt. Hierdoor kan de Virtuele Verhalenverteller tekst creëren die gemarkeerd is met een emotie. Deze tekst wordt daarna omgezet naar emotionele spraak.

Abstract

The development of virtual story-telling is an ever ongoing process. As long as it does not perform on the same level as a human storyteller, there is room for improvement. The Virtual Storyteller, a project of the HMI department of the University of Twente, uses a Text-to-Speech application that creates synthetic speech from text input.

It is definitely not a trivial matter to make synthetic speech sound human, especially when story-telling is involved. There are so much facets of story-telling where storytellers use their voice in order to enhance their performance. This thesis describes the study of how storytellers use their voice in order to convey the emotions that characters in the story are experiencing; specifically how the storyteller changes his voice to make a character say something in an emotional way.

An experiment has been conducted to identify the emotional charge in fragments of speech by story-characters. These fragments are then analysed in an attempt to find out how the emotional charge is linked to the way the storyteller changes his voice. The analysis is then used in the creation of a model that is used by the Text-to-Speech application to synthesise emotional speech instead of neutral speech.

This model is implemented in an open-source Text-to-Speech application that uses a Dutch voice. This allows the Virtual Storyteller to create tekst marked with an emotion, which is then used to synthesise emotional speech.

Contents

Chapter 1

Introduction

This chapter will give a quick introduction into story-telling, the virtual story-telling project, and how to enhance the performance of story-telling using a TTS engine.

1.1 Story-telling, what is it and why do we do it?

Story-telling is a very multifaceted subject, not consistently defined and very open to subjective interpretation. A lot of people define story-telling in a different way, but in my opinion, the National Story-telling Association [16] uses the best refined definition.

> Story-telling is the art of using language, vocalisation, and/or physical movement and gesture to reveal the elements of a story to a specific, live audience.

These story elements are informational in nature, and vary between simple absolute facts (consistent with the world we live in or not), reports of facts by characters, which makes the reports subjective and therefore not necessarily true, information on a higher level which consists of not only facts and reports, but also guidelines on how to proceed in possible unknown situations (like how to fish), wisdom, which involves morals and ethics, and feelings which can include emotions [6]. As can clearly be seen, stories involve passing on information, on any level.

There are many reasons why one would tell stories. We will call these story-telling goals: to educate others in a historical-, cultural- [50], or ethical aspect, entertainment, or for personal development such as linguistic- (vocabulary and listening skills), or psychological (self-awareness and world comprehension) development [18][8]. If a story is told for entertainment purposes, such a story will always contain one (or more) of the other story-telling goals because of the story elements present in the story. If a story is about nothing, it is very likely that it is not entertaining and thus the story has to contain story elements which end up as a part of the other story-telling goals. The reverse is also true: Entertainment is also a requirement for the successful transfer of the story contents. If the story is not entertaining, people will start to lose attention and part of the story elements, and with it part of the story goals, will be missed. As such, entertainment is both a requirement and a result of a successful story-telling act. Some people make the distinction between story-telling performed with a one-way information stream, such as listening to or watching prerecorded audio, video or text, and a two-way information stream such as live performances of storytellers, with the only distinction being that the performer can adapt his/her performance depending on the feedback given by the audience, which can be anything from noticing people are not paying attention to the storyteller to massive jeering.

1.2 Virtual story-telling

At the University of Twente, research is being done to create a virtual storyteller[1] [54]. This project involves creating the plot of a story, creating the language to express the story, and presenting the story to the audience. The VST system performs the roles of both the author and the narrator. The characters of the story are represented by intelligent emotional agents which, guided by the plot agent where necessary, reside in a world managed by a world agent. The actions of these agents are monitored and, after additional language processing[47], transformed into a story. This story is then presented to the audience by means of a Text-to-Speech application. The internal emotional state of the agents however, is only used to influence the actions of the agent, and is not used at all to enhance the presentation of the story. The VST system currently only generates a narrated story, without any character dialogue. Most fairy-tales however, do contain character dialogue and it is not unexpected that future versions of the VST system will also create character dialogue. At the moment, the VST system is unaware of any audience, and is unable to get (and thus use) feedback from the audience. The current naturalness of the story-telling voice used in the VST system is nowhere near a real life storyteller. It depends on already defined rules of accent placement and pronunciation which are not designed for a story-telling performance and will therefor need updating to conform with the style a professional storyteller uses. Also the VST does not use emotions in the narration of a story.

1.3 Goals

This project focuses on how a given story can be told best by relying on how popular storytellers tell their stories. In particular, the aspects of story-telling related to the Text-to-Speech (TTS) system will be looked at: how to make your voice sound when telling a story. The wording of the story itself is predetermined, the focus lies solely on presenting the text to the audience in a most lifelike manner. The ultimate target performance of the TTS system is to mimic a real storyteller, voice actor or just someone who is good at reading aloud. This means that the presentation of the story-elements must be done transparent to the story-telling act itself; people should not notice how good the storyteller is, but should be completely immersed in the story [51].

The current TTS implementation of the VST system uses a closed-source commercial engine. This limits the amount of influence one can exert in order to change the way the speech is generated. At the moment the process which augments the voice of the TTS engine with story-telling styles takes place outside the engine because of the inability to influence the necessary prosodic parameters inside it [27]. As such, one of the goals is to implement this process using an open-source TTS engine in which the amount of influence that can be exerted to the parameters needed for the generation of speech is greater than it is in the current situation.

I will investigate if the generated speech can be improved to sound more like a natural storyteller. In order to allow a TTS to mimic a good storyteller, it is important that the material used (instructions or audio) to base measurements on comes from a generally as 'good' accepted storyteller or voice artist. Koen Meijs has performed an in-depth investigation about the prosodic configurations of the narrative story-telling style and two types of usage of suspense in story-telling and devised a rule set that allows the morphing of a neutral voice in order to create suspense and emphasize words in a way storytellers do [27]. This, combined with the fact that the emotional state of the agents in the VST system is currently unused with respect to the narration of the story, is the reason that I will focus on the emotion/attitude aspect of the story-telling act.

Initial literature study on emotions and story-telling showed that the subject of emotions and story-telling is too vast to be completely analysed in this thesis. Because of this, I have chosen to limit my research to the study of emotional utterances of characters inside a story and how to recreate these utterances using TTS. The end result is a number of software applications which make the transformation from emotion-tagged text (produced by the VST story generator) to emo-

[1] The Virtual Storyteller will henceforth be abbreviated with VST

tional synthetic speech possible. It also includes the narrative styles that have been implemented by Meijs in an earlier version of the narration module of the VST.

1.4 Approach

First, literature has been consulted in order to determine how emotions are used in the art of story-telling. The results of this are given in section 2.1. Next, the reader is supplied with a brief overview of current TTS technologies in section 2.2. Also, a brief literature study on prosody and how it is used in this thesis is presented in section 2.3, followed by a short overview on how emotions are categorised in section 2.4.

In chapter 3, several TTS engines are listed and the choice is made on which TTS engine will be used in the remainder of this project.

It quickly became apparent that when you want to study aspects of fragments of emotional speech, you must first annotate these fragments with emotion terms. In order to do this, a perception test was created. In this experiment, fragments of a told story are linked with the emotion perceived. This is reported in chapter 4. The fragments which scored well above chance were deemed eligible for detailed analysis in order to determine which prosodic parameters were largely responsible for the perceived emotion. The analysis and its results as well as the first model created are detailed in chapter 5. However the analysis was not conclusive and the model was discarded in favour of another model. This other model is presented in chapter 6.

After this, the new model has been implemented in the chosen TTS system. In order to interface the new model with the VST system, an xml format was devised. The TTS system has been extended with the ability to use SSML (see section 2.2.4) as input. A preprocessor application was designed and implemented that converts the VST xml files into SSML. All this is documented in chapter 7.

In order to confirm (or deny) the proper functioning of the new model, an evaluation test was performed. Test subjects were presented with two variants of an utterance: one variant uses the developed new model to change the prosody, the other variant does not. The test subjects were then asked if they recognised the target emotion in the modified version. The test and its results are shown in chapter 8.

Chapter 2

Literature

This chapter contains the different contributing factors of story-telling, gives an overview of existing speech synthesis technologies, set a definition for 'prosody', and show the emotion classification used further in this project.

2.1 Story-telling

Even though the term is widely known, and lots of people have practised or witnessed acts of story-telling themselves, there are a few things that need to be distinguished in order to have some specific subjects to focus the research on.

In the life-cycle of a story, three different parties can be distinguished:

- The author: the author creates the story world, people, and events within the story.
- The narrator: the narrator presents (tells, shows) the story to the reader.
- The reader (or audience): the reader's contribution is to understand and interpret the story.

The VST performs the roles of the author and the narrator, creating the story from scratch and presenting it to whoever is present. This research will focus on the narrator, in order to improve the VST narrator's naturalness. As such, it is good to know that when a narrator uses vocalisation in order to tell a story, several different aspects of story-telling techniques distinguish themselves [27] page 10,11:

- Specific narrative style
  A narrator who tells a story uses a completely different speaking style than a news reader. 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 stories for children.
- Presence of emotion and attitude
  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 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).

With the exception of character voices, Meijs [27] has modelled the above mentioned aspects of the narrative story-telling style. Since the goal of this thesis is to enhance the VST system with the ability to produce emotional expressions during the speech synthesis, we must ask ourselves where and why, in a story-telling act, emotional expressions occur.

There are multiple reasons why an emotional charge can be present in a fragment of a story, for example:

- The storyteller is (accidentally or not) experiencing an emotion while telling the story which influences his performance
- The storyteller wants to make a character of the story sound like it's experiencing an emotion, which he implements by either recalling an emotional event and thus experiencing a target emotion which influences his performance, or by faking an emotion.
- The storyteller wants the audience to experience a specific emotion

In a paper by Véronique Bralé et al. [3], who investigate the expressiveness in story-telling, she states (pg. 859):

> Expressiveness is used here to refer to voluntary communicative affective expression as opposed to involuntary and direct expression of the speaker's emotional states during his/her speech communication process. Of course, in the specific case of story-telling, we are not interested in the speaker's own emotional states, but either in the expression of the emotions felt by the story characters or in the expression aiming at inducing emotional states to the listener.

I wholeheartedly agree with this. While involuntary expressed emotions can influence a storyteller's performance, TTS systems of today do not have emotions of their own and therefore research should be directed towards the voluntary expressed emotions. Furthermore, there is a difference between emotional expressions and expressions intended to cause emotions. When analysing the expressions themselves, it is easier to recognise the former, but not so with the latter: almost all human beings can determine if an expression is emotional -you can determine if someone sounds afraid, angry, bored etc- but figuring out if an expression is intended to cause emotions just by analysing the expression is not as easy (except when this is done by expressing the emotion of the character: people might start to feel affection for a scared little character). The cues of expressed emotions experienced by in-story characters are also present in the expression itself (and the context in which the expression is performed), which is not necessarily the case when inducing emotional states in the listener. Inducing an emotional state could work or not, there is no telling if it is effective on the entire audience just by analysing the speech. Because this thesis' goals focus on improving the quality of a TTS system by analysing speech of professional storytellers, I will focus on the expressions aimed at showing emotions, not inducing them. Emotions of characters can be communicated indirectly by the speaker by describing a character, or directly, by allowing the characters to express their own emotions. The most obvious method of emotion expression that is used in story-telling is a character utterance. A character might scream: "Run...run for your lives!" in fear (and thus in a fearful voice); this is a direct expression of emotion.

The VST system currently does not feature character dialogue, but since character dialogue is used in fairy-tales, I expect it will be included at some point in the future. It does however employs emotional agents to construct a story, but, at the moment, uses the emotional states of the agents only to motivate their decisions. These emotional states can therefore also be used in the narration of the story by changing the prosody[1] of the speech produced for characters by the TTS system. With this, the VST system can be extended to allow the characters to sound emotional. That is, once character dialogue has been added.

There is another aspect of character utterances that is used by professional storytellers which deserves special attention because it complicates matters when using a TTS system to perform the role of a narrator: These are the non-linguistic vocal effects such as laughing, giggling, tremulousness, sobbing and crying. They cannot be generated using a diphone database (see section 2.2.1 for an explanation of what this is). It would be possible to record a database with samples of these effects performed by the same speaker that was used to record the diphone database, but

[1] More on prosody in section 2.3

besides the effort of doing so there are a few other problems that arise when intending to use this collection of vocal effects:

- How do we know when to use which sample; what distinguishes one giggle from another?
- How many different recordings of each effect would we need in order for the audience to not notice the fact that the recordings are reused?
- How well can these recordings be inserted into the generated speech audio-stream without unwantingly disrupting the continuity?

There has been little or no research into different forms of these effects and their prosodic configurations, like why one giggle is different from the other and how this can be measured. This makes it impossible to use them in TTS systems without extensive additional research.

And lastly, there is the use of voices for characters. Although its use is quite influential, It is my observation that storytellers (knowingly or not) modify their voice by changing spectral parameters (see table 2.1, which is at the moment not a freely modifiable parameter in most TTS systems. To my knowledge, there is no research done on the specific voice changes when using character voices.

2.2 An overview of speech synthesis

This section gives a quick glance on the various types of speech synthesis, based on the works of [13]and [26].

The production of human speech by means of machines is called speech synthesis. This can be done with specialised machines, which are built specifically for this purpose, or general purpose machines which can be configured to perform the speech synthesis, like computers. These systems are usually referred to as Text-To-Speech (TTS) systems, as the input of such a system consists of text, and the output consists of produced human speech.[2]

The function of a TTS system can be divided into two main parts: a front-end and a back-end part. Since the input of a TTS system is in text form, this text will need to be interpreted and ambiguities will have to be removed. This is the task of the front-end part. The output of this, some form of representation about what is to be 'said' exactly, is the input of the back-end system, which converts it to the human speech output.

The interpretation of the input text consists of converting numbers and abbreviations to the text that people say when they read items like that. For example, the input '6' will be translated to 'six' and '500' to 'five hundred'. Things like dates, abbreviations, and phone numbers (which you don't want to interpret as being one big number) are changed to their full written-out form by the front-end system. After this, phonetic transcriptions will be made for the entire text. The whole text will also be chunked into sentences, phrases and prosodic units. After this, the data goes to the back-end system.

The back-end system converts the preprocessed data directly to audio (be it directly audible, or a playable waveform datafile). This can be done in several ways: Utilising concatenative synthesis (unit selection synthesis, diphone synthesis, or domain-specific synthesis), formant synthesis or articulatory synthesis.

2.2.1 Concatenative synthesis

The concatenation of pieces of previously recorded speech is at the basis of concatenative synthesis. These pieces can be anything from something as small as a single phone (sound) up to complete sentences. All these pieces are categorised by certain acoustic correlates (like fundamental frequency, duration, phones directly in front and behind, position in the syllable, etc) and put in databases. When any input needs to be converted to speech, each bit of input will be checked and the best matching unit in the database will be selected to be used. This method is used for the 'unit selection synthesis' method.

[2] For a full and in-depth overview of all speech synthesis methods and history, reading [26] is highly recommended.

The diphone synthesis method uses diphones (phone-to-phone transitions), which were extracted from prerecorded speech. Each language has a different diphone set because there are certain transitions that don't occur in one language that do happen in another, and different phone sets as well. For the speech synthesis, these diphones are concatenated, and the resulting audio-data is then modified using digital signal processing methods like PSOLA, MBROLA and LPC (Linear Predictive Coding). A comparative research project on these algorithms was done by Dutoit[13].

Domain-specific synthesis works on a slightly higher level, in such that it concatenates whole pre-recorded words and phrases together, instead of smaller units or diphones. This results in quite natural sounding synthesis, but greatly limits the different amount of sentences that can be synthesised without needing an exceptionally huge database.

2.2.2 Formant synthesis

Formant synthesis is based on acoustic models instead of pre-recorded human speech. These models contain rules on how to use fundamental frequency, volume, voicing, noise and other acoustic correlates over time in order to create speech. Because of this, no large databases with human speech recordings are required, allowing this method to be applied in areas where memory is a limited resource. Also because of the fact that no human speech recordings are used, the speech synthesised is not bound by the qualities of the voice used. This way the system is not bound by the ability to output only one voice, but can create different voices in different moods changing more than just the fundamental frequency and tempo, but also spectral parameters, breathiness, creakiness and more. On the downside, the synthesised speech sounds a lot less natural than the speech synthesised with concatenative methods.

2.2.3 Articulatory synthesis

The method that tries to synthesise human speech while staying close to the way humans generate speech is articulatory synthesis. With this, models are made of the function of the entire human vocal tract in order to generate speech just as we do. Needless to say, these models and their functions are extremely complex, and few synthesisers are actually able to employ this method sufficiently.

2.2.4 Speech synthesis markup languages

There exist a few different formats besides raw text to provide speech synthesisers with input. These formats allow for extra data to be supplied to the synthesiser such as changes in speed, volume, frequency. Among these are SABLE[44] which claims to have been expanded upon SSML[49]. But SSML was later improved based upon what SABLE had become. There is a proposed extension to SSML that includes higher-level prosodic parameters such as ToBI labels and abstract concepts like emotions[14]. Another format is JSML[24], based on the Java programming language. Although each of these was proposed as a new standard, still none of them has been widely adopted.

2.3 Prosody

As stated earlier, the methods for telling a story are through language, vocalisation and/or physical movement and gesture. A goal of this project is to increase the naturalness of a TTS-system's story-telling capabilities, and since a TTS-system is unable to perform physical movements and gestures, those methods are ignored in this project. The text of the story is predetermined, created by another part of the VST system, therefore the language aspect is also not included from the scope of this project, leaving us with only the vocalisation. This vocalisation is explained using the terms prosody and paralinguistic features below.

As with the definition of story-telling, the definitions of prosody and paralinguistic functions vary depending on who you ask. The Merriam-Webster dictionary defines prosody as "the rhythmic and intonational aspect of language". When looking up paralanguage, which Merriam-Webster defines as "optional vocal effects (as tone of voice) that accompany or modify the phonemes of an utterance and that may communicate meaning", it becomes clear right away that there is some overlap with prosody.

Schötz [45] compared all the various definitions and observations. While not resulting in clear definitions for both prosody and paralinguistics, it does shed some light in this matter. Firstly, she shows that there are multiple perspectives for observing the various properties of prosody: On an acoustic level, on a human perceptual level and on a functional level. The first two are linked in such a way, that the data is correlated, but the terms used to express them differ. For example: What's measured as the 'amplitude' of a (speech) signal in the acoustic level is called the 'loudness' of the voice in the human perceptual level. The functional level describes what each property is used for. Since most functions are not defined by specifically one property, but by a combination of properties, there is a bit of overlap here. The functions are realized by certain configurations and variations in the perceptual correlates, which are in turn realized by different configurations and variations of the acoustic correlates. Table 2.1 has been copied from [45], and shows the properties of prosody as mentioned in the literature on prosody she researched. These properties also contribute to the paralinguistic functions like the signalling of sex, age, physical state, emotion and attitude. This table introduces some linguistic lingo which are explained here. F0, also known as fundamental frequency, is used to indicate the first the base frequency of a speech signal. It is also the first (and most noticeable) formant. Formants are certain frequencies which are more prominent than others in the human voice. A human male has 5 formants, and a human female 4. These formants are formed by the resonance properties of the oral- and nasal cavity, as well as the pharynx. The F0 correlates with what we perceive as the pitch of someone's voice (which can be seen in table 2.1. The F0 contour is the measured behaviour of the F0 over time. The F0 range is the domain of frequencies over which the F0 varies. The F0 level is the absolute minimum F0 observed in an utterance. The voice quality is mostly measured in terms of breathiness and creakiness. Whispering has a very high breathiness quality, while if you try to talk in a really low voice, the creakiness quality becomes quite apparent. The distinctive function of prosody allows us to distinguish between words, that are spelled the same but mean something different, by using variations in prosody. Formant undershoot happens when trying to talk faster. It is indicated by a reduction in the amount of effort being taken to utter vowels. For example when saying "iai" repeatedly, you will notice that the amount of effort (and with it the purity of the vowel) decreases as you go faster. Paralinguistic functions are used to signal properties such as: speaker age, sex, physical condition, emotions and attitude.

Each utterance made by man expresses not just the words that were said, but also how they were said. This is influenced by the physical and emotional state of the person who uttered it and the informational structure, and is called the prosody of the utterance. In this report, the term 'prosody' will mean exactly that: Everything about an utterance that are not the words themselves. The paralinguistic functions that give information about the physical and emotional state of a person, like gender, age, anger, happiness, illness, are a subset of the prosody of an utterance. Hence, the emotional aspects of an utterance that will be researched in this report will be labelled under prosody.

2.4 Definitions of emotions

As with defining prosody, classifying human emotions is a subject on which a lot of people have quite an amount of different views and opinions. As mentioned in [31], there are several people who have created lists of different emotional terms: [23][38][15][19], but there is no single list or categorisation of emotions that is commonly accepted. The tree-structured list of emotions, as described in [37] (see table 2.2) will be used in the remainder of this project. This table is extensive enough to accommodate most terms used by everybody else, and yet structured in order to be

Properties	**Melody**	**Rhythm**	**Dynamics**	**Spectrum**
Acoustic correlates	F0 (contour, range, level, movements)	Duration (of segments, syllables including silence), quantity	Intensity (amplitude)	Phonation type, formant frequencies
Perceptual correlates	Pitch (contour, range, level, movements), intonation, tone	Length, speech rate (tempo), rhythmicality?	Loudness	Segmental quality, voice quality
Functions	Distinctive (lexical tone, stress), prominence (stress, accent, focus), structural (grouping, boundary, sentence type, speaking style), paralinguistic functions	Prominence (stress, accent, focus), structural (grouping, boundary, sentence type, speaking style), paralinguistic functions	Prominence (stress, accent, focus), structural (grouping, boundary, sentence type, speaking style), paralinguistic functions	Less reductions (formant undershoot) in stressed syllables, paralinguistic functions

Table 2.1: Functions, Acoustic and Perceptual correlates of prosody.

able to derive the primary emotion for each of the used terms. The tree-structure of the table is a representation of the division of emotional terms into subsets. Primary emotions are terms of which the secondary emotions are a (more specific) subset, and similarly secondary emotions are terms of which the tertiary emotions are a (more specific) subset.

How prosody is used to signal emotion, and how this can be analysed - by showing which acoustic correlates play a role in this process - is dealt with in chapter 5, specifically sections 5.1 and 5.3

Primary emotion	**Secondary emotion**	**Tertiary emotions**
Love	Affection	Adoration, affection, love, fondness, liking, attraction, caring, tenderness, compassion, sentimentality
	Lust	Arousal, desire, lust, passion, infatuation
	Longing	Longing
Joy	Cheerfulness	Amusement, bliss, cheerfulness, gaiety, glee, jolliness, joviality, joy, delight, enjoyment, gladness, happiness, jubilation, elation, satisfaction, ecstasy, euphoria
	Zest	Enthusiasm, zeal, zest, excitement, thrill, exhilaration
	Contentment	Contentment, pleasure
	Pride	Pride, triumph
	Optimism	Eagerness, hope, optimism
	Enthrallment	Enthrallment, rapture
	Relief	Relief
Surprise	Surprise	Amazement, surprise, astonishment
Anger	Irritation	Aggravation, irritation, agitation, annoyance, grouchiness, grumpiness
	Exasperation	Exasperation, frustration
	Rage	Anger, rage, outrage, fury, wrath, hostility, ferocity, bitterness, hate, loathing, scorn, spite, vengefulness, dislike, resentment
	Disgust	Disgust, revulsion, contempt
	Envy	Envy, jealousy
	Torment	Torment
Sadness	Suffering	Agony, suffering, hurt, anguish
	Sadness	Depression, despair, hopelessness, gloom, glumness, sadness, unhappiness, grief, sorrow, woe, misery, melancholy
	Disappointment	Dismay, disappointment, displeasure
	Shame	Guilt, shame, regret, remorse
	Neglect	Alienation, isolation, neglect, loneliness, rejection, homesickness, defeat, dejection, insecurity, embarrassment, humiliation, insult
	Sympathy	Pity, sympathy
Fear	Horror	Alarm, shock, fear, fright, horror, terror, panic, hysteria, mortification
	Nervousness	Anxiety, nervousness, tenseness, uneasiness, apprehension, worry, distress, dread

Table 2.2: Emotions categorised by Parrott [37]

Chapter 3

Choosing the voice

This chapter will explain which TTS system is selected to be used and detail the features of the chosen TTS system.

3.1 Text-to-Speech engine

As stated earlier, we would like to use an engine in which the amount of influence that can be exerted to any aspects of the speech-generation process is as large as possible. In order to do this, an open-source TTS engine will immediately be at an advantage because you will be able to investigate the entire engine at the code level, modify existing code to do the precise things you want it to do, and add code in case you want to extend the engine's current functionality to include something new. In addition, because the VST system generates stories using the Dutch language, the engine will have to support the generation of Dutch speech. There are a few open source TTS systems available:

- The Festival Speech Synthesis System[1]
  Festival offers a general framework for building speech synthesis systems as well as including examples of various modules. As a whole it offers full text to speech through a number APIs: from shell level, through a Scheme command interpreter, as a C++ library, from Java, and an Emacs interface. Festival is multi-lingual (currently English (British and American), and Spanish) though English is the most advanced. Other groups release new languages for the system. And full tools and documentation to build new voices are available through Carnegie Mellon's FestVox project.

- Flite: a small, fast run time synthesis engine[2]
  Flite (festival-lite) is a small, fast run-time synthesis engine developed at CMU and primarily designed for small embedded machines and/or large servers. Flite is designed as an alternative synthesis engine to Festival for voices built using the FestVox suite of voice building tools.

- FreeTTS[3]
  FreeTTS is a speech synthesis system written entirely in the JavaTM programming language. It is based upon Flite: a small run-time speech synthesis engine developed at Carnegie Mellon University.

- GNUspeech[4]
  Gnuspeech is an extensible, text-to-speech package, based on real-time, articulatory, speech-synthesis-by-rules. That is, it converts text strings into phonetic descriptions, aided by a

[1] http://www.cstr.ed.ac.uk/projects/festival/
[2] http://www.speech.cs.cmu.edu/flite/
[3] http://freetts.sourceforge.net/docs/index.php
[4] http://www.gnu.org/software/gnuspeech/

pronouncing dictionary, letter-to-sound rules, rhythm and intonation models; transforms the phonetic descriptions into parameters for a low-level articulatory synthesiser; and uses these to drive an articulatory model of the human vocal tract producing an output suitable for the normal sound output devices used by GNU/Linux.

- The Epos Speech Synthesis System[5]
  Epos is a language independent rule-driven Text-to-Speech (TTS) system primarily designed to serve as a research tool. Epos is (or tries to be) independent of the language processed, linguistic description method, and computing environment.

- MARY[6]
  MARY is a Text-to-Speech Synthesis System for German, English and Tibetan. It was originally developed as a collaborative project of DFKI's language technology lab and the Institute of Phonetics at Saarland University and is now being maintained by DFKI. MARY is designed to be highly modular, with a special focus on transparency and accessibility of intermediate processing steps. This makes it a suitable tool for Research and Development.

Of the above, there only appears to be a (maintained) Dutch language and voice module for Festival, namely NeXTeNS[7] which stands for "Nederlandse Extensie voor Text naar Spraak" (Dutch extension for TTS). Although it would be possible to create a Dutch front-end (language parser) and back-end (voice) for GNUspeech, Epos or Mary, this would create a significant amount of extra work. Therefore, the Festival system, together with the Dutch extension NeXTeNS, is the TTS system that we will be using.

3.2 Limitations of Festival

Festival, using a diphone concatenation MBROLA engine, uses a fixed database of prerecorded diphones uttered by one speaker (per database). The TTS engine allows modification of all the attributes that fall under the category of 'Acoustic correlates', with the exception of those under 'Spectrum' and as such should be able to achieve expressiveness in the attributes listed next to 'Perceptual correlates' and 'Functions'; once again except those under 'Spectrum'. Because of the use of diphone concatenation, the voice qualities, as listed under 'Spectrum', such as: modal voice, falsetto, whisper, creak, harshness and breathiness cannot be changed easily. This is caused by the fact that diphone concatenation uses a database of fixed diphones previously recorded from a single speaker, and thus all recorded diphones from that database will have the same voice quality (as listed under 'Spectrum'). If we want to vary the voice quality on-the-fly, the only method to achieve this is to load a different diphone database recorded from the same speaker but under different (emotional or physical) circumstances. Still, if these databases were available (in the proper format) the change of the 'overall' voice might be too noticeable and the loading time might seriously hamper the continuity of the produced speech output.

3.3 Conclusion

The condition that the TTS platform to be used must be open-source and have a Dutch front- and back-end quickly limits the available options to one: Festival. This system is designed to be used as a research platform and is therefore not the fastest or the most memory-efficient TTS system available, but it gets the job done. The limitations described in the previous section are a result of the use of a diphone concatenation engine. This back-end engine is used in most TTS systems as it gives good results and takes (relatively) little effort. It would be possible to use a different technique, however such a technique would have to be implemented first, and creating a Dutch TTS back-end from scratch is not the objective of this thesis.

[5] http://epos.ure.cas.cz/

[6] http://mary.dfki.de/

[7] http://nextens.uvt.nl/

Chapter 4

The search for emotion

In this chapter, the setup and results of a series of experiments is presented. The experiments have been designed to test for the presence of - and if present, identify - the emotional charge in fragments of speech from a story told by a well known storyteller. These fragments are then analysed in terms of acoustic correlates. The resulting measurements are used to construct a model that links the various identified emotions to these acoustic correlates in order to update a TTS system to produce emotional sounding utterances.

As deduced earlier in section 2.1, we are interested in the direct expression of emotions by characters in a story. In order to obtain audio with emotional utterances of in-story characters prerecorded fairy-tales were scanned for fragments consisting of utterances spoken by characters inside the story. An example of a short, utterance which can contain an emotional charge is this: ”Who goes there?”. Obviously, figuring out which emotion is experienced by just looking at this bit of text is rather difficult. When given the context, it becomes much clearer: ”As he hesitantly sneaked down the corridor towards the strange sound he was hearing Jan almost jumped against the ceiling when he suddenly felt something lightly touch his shoulder. ’W..Who goes there?’”. Even though the utterance itself is short it clearly shows that there should be a difference between how the narrative part is pronounced and how Jan’s question is pronounced. In most cases, stories also describe how utterances are pronounced by adding descriptions like “, he said” or “, he yelled” or even more detailed like “, he said terrified”. These added descriptions are usable to determine which emotion is to be used when reading a story aloud. Back to the example, Jan is (obviously) scared out of his wits and thus his question is likely spoken with a matching frightened voice. Because the emotional charge of the utterance by Jan is so obviously present, the prosodic attributes specific to this emotion should be easy to detect when analysing an audio fragment with this utterance.

When looking at the utterances by in-story characters, and matching the perceived emotions with the emotional state we believe the character should be experiencing (from context), we can identify the utterances which have been uttered efficiently (as in, the perceived emotional charge of the utterance can be associated with the believed emotional state of the character) and use those to create a model to allow a TTS system to express emotions experienced by in-story characters.

It is possible that utterances do not contain merely one single emotion in the emotional charge, but a mixture of several emotions. For example, the emotional charge of an utterance of someone who is “pleasantly surprised” will likely contain the emotions (or influences thereof) happiness and surprise, and as such, the utterance will probably display prosodic attributes for both these emotions. This makes identifying the prosodic attributes of a sole emotion more difficult compared to when only one emotion is present. Assuming that the more prominent an emotion is in the (possible) mixture of emotions in an emotional charge, the more prominent its prosodic attributes will be, then the utterance can be categorised as having the most prominent emotion as ’the’ emotion of that utterance without influencing the impact of the measured prosodic attributes (because the most prominent prosodic attributes attribute to ’the’ emotion).

Besides being able to link various emotions to certain story fragments, a baseline is needed which consists of fragments which are considered to have no emotional charge (emotionally 'neutral'). This means that the test-data for the experiment also has to contain fragments of which the result will be that people will find there to be no emotional charge present. This will allow us to directly compare emotionally neutral utterances to emotionally charged utterances. If there are no 'no emotional charge present'-fragments for a certain character, the emotional neutral baseline will be missing, and a model specifying how a voice is changed when comparing 'no emotion' with another emotion (any) cannot be setup from those fragments.

This chapter is set up as follows: In section 4.1 the constraints are described for the audio material that is to be analysed. Then the interface of the experiment and the degree of freedom of the participants in answering the questions of the experiment are investigated in a preliminary experiment, described in section 4.2. After that, the constraints on the participants and the environment in which the experiment has taken place are specified in sections 4.3 and 4.4. The interface determining what a participant gets as input and what he should answer are finalised in section 4.5 and finally the results of the experiment are discussed in section 4.6.

4.1 Test material

Because the aim of this research is to mimic professional storytellers using TTS, the material used for the experiments comes from professional storytellers. There is a large amount of told fairy-tales available on tape, for example the "Lecturama Luistersprookjes", which were commercially available in 1986. These stories were a huge success, and thus it can be concluded that the quality of the storytellers is generally agreed upon to be good. This qualifies the stories to be used as a basis for extracting features of story-telling that is generally accepted as being of good quality. The fairy-tales have another feature which makes them more interesting for this test: they are child-oriented. This means that the narrators and authors employ strong (exaggerated) and simple methods of conveying the story contents so the children will pick it up more easily. Sentences and words are short and simple, accents and emotions are emphasised. In order to be able to morph a TTS voice from sounding neutral (which is the default) to sounding emotionally, both neutrally- and emotionally sounding fragments of in-story character dialogue from fairy-tales are needed. However, there is a risk that all of the fragments of a certain character will be labelled with the same emotion (for example, happy), just because the character sounds happy all the time. This would result in unusable fragments, because without a baseline 'neutral' fragment, there is no way of knowing which specific acoustic correlates make the character sound happy. The only useful information, if other fragments by the same character are labelled with a different emotion, would be how to change the acoustic correlates to make a happy voice sound, for example, sad. Since TTS systems do not employ voices that sound happy by default, this information is rather useless and the entire set of fragments of that character cannot be used to change an emotionally neutral TTS voice to one that contains an emotional charge.

The test material was selected from the audio available of the Lecturama Luistersprookjes with the following criteria:

- The storyteller must be the same throughout all the fragments.
- The fragments must contain as little as possible different characters.
- The syntactical part of the fragments must preferably not be biased towards a certain emotion.

The reasoning behind these criteria is as follows: A single storyteller means that only one voice is used for all fragments. This means that all biological-specific attributes that determine the voice quality (like gender and age) are the same throughout the entire set of fragments. Similarly, having the most fragments from the least amount of different in-story characters means that the prosodic configuration between all the fragments of one character are similar. After all, it is the 'voice' of that specific character. In order to avoid giving information away that could give the test user an indication of the emotional charge present in the fragment from anything other than the vocalisation of the fragment, (for example: "Oh no...help me...help me!" which indicates distress) fragments of which the text leads to a clear choice of which emotion is present are excluded. This is done in order to make sure that test users base their choice (of which (if any) emotion is present in the fragment) on the prosodic aspects of the utterance as much as possible.

With this in mind, 52 stories were analysed. The stories used in the preliminary tests (see below) were excluded from the final material. Of the remaining stories, three remained that contained enough material to work with. One of those three stories consisted of a dialogue between two people, and resulted in two times seven usable fragments. The other two stories yielded one set of fragments each, one consisting of four and the other of 12 fragments.

As mentioned in the beginning of this chapter, the context can show which emotion someone is currently experiencing. To this end, the fragments which have been selected for the presence test have been analysed, and the emotional state derived from the context has been written down. This is done by looking for explicit statements about the characters emotional state, and by looking at the description on how a character speaks. Also taken into account are the motivations of the characters. For example, Magda is being secretive/protective about a received letter, and so when she tells her husband to close the door, she might do so in a slightly fearful voice. The entire analysis consists of quite an extensive amount of text, and therefore has been included in the appendix as table A.6. The fragments there are presented in chronological order. A quick situational update has been constructed by me and put between parentheses (). The character quote which has been used in the presence test has been emphasised; the rest of the text is spoken by the narrator. The emotions from this context analysis are compared to the results of the presence test later on.

4.2 Preliminary experiment

The Lecturama Luistersprookjes feature two different storytellers, one male and one female[1]. As such, two experiments were performed, one with fragments told by the male storyteller and one with fragments told by the female storyteller, in order to determine which one would be the best to use in the final experiment. These two experiments were performed by 3 people, of which two agreed that the female voice sounded better. The third found the male voice better, because the female voice was, in his opinion, too theatrical. But theatrical means that the prosodic cues are presented more explicitly, and therefore easier to pick up on. This is exactly what we wanted and thus the stories told by the female voice were used in the final experiment.

In order to determine if an open answering mechanism was the best way for the user to give his answers, as well as to determine which emotions should be presented as a possible answer, another experiment was performed. In this experiment, which was performed by three people knowledgeable in the field of linguistics and myself, the question "Which emotion do you hear?" was an open question. The experiment consisted of 27 fragments. This resulted in people entering whatever term they thought of (which was not necessarily an established term for an emotion). After the experiment was performed, all answers were analysed, and identical answers were counted. Since numerous terms were entered which did not directly indicate an emotion, some were 'rounded' to the nearest applicable emotion using table 2.2, which shows a list of primary, secondary and tertiary emotions originally presented in [37]. The different terms used as answers, the frequency of the term usage, and the reasoning behind the rounding to the nearest emotion are shown in table 4.1.

The diversity of the answers, as well as the number of different terms used, were - as far as the goals of this preliminary experiment were concerned - a clear indication that multiple-choice answers were needed instead of open answers. This is to steer the participants of the test away from using terms that are not really emotions in the first place, which would result in a lot of different terms that occur with a low frequency which decreases the importance of each individual term. If thirty people each use a different term for one sentence, no conclusive result can be ascertained about the significance of any singular term. Because we want to be able to single out one emotion

[1] Although no explicit names are present on the cover of these cassettes, they are probably Frans van Dusschoten and either Trudy Libosan or Martine Bijl

Term used (Dutch)	**Times used**	**Emotion derived**
Angstig	4	Fear
Belerend	1	Unknown
Berustend	4	Resigned
Beteuterd	1	Unknown
Blij verrast	1	Surprise
Blijheid	8	Joy
Boosheid	2	Anger
Droevig	3	Sadness
Enthousiast	1	Enthusiasm - Zest - Joy
Gedreven	1	Passionate/enthusiasm - Zest - Joy
Gelaten	1	Unknown
Geruststellend	3	Caring - Affection - Love
Irritatie	1	Irritation - Anger
Laatdunkend	1	Unknown
Leedvermaak	5	Malicious Delight - Delight - Cheerfulness - Joy
Lijdzaam	1	Unknown
Medelijden	5	Compassion - Affection - Love
Minachtend	1	Contempt - Disgust - Anger
None	41	None
Opgelucht	1	Relief - Joy
Opgetogen	5	Delight - Joy
Opgewekt	1	Cheerfulness - Joy
Opgewonden	1	Excitement - Zest - Joy
Standvastheid	1	Unknown
Verheugd	3	Gladness - Cheerfulness - Joy
Verrast	1	Surprise
Verruktheid	1	Delight - Cheerfulness - Joy
Verwachting	1	Unknown
Verwijtend	2	Unknown
Verzekering	1	Unknown
Verzuchtend	1	Unknown
Wanhoop	1	Despair - Sadness
Zeurend	1	Unknown

Table 4.1: emotion modifications in the preliminary experiment

and link it to a fragments as being 'the' emotion of the emotional charge of that fragment, the use of a large number of terms should be avoided. To be on the safe side, an 'Other'-option was still left in for users to enter a custom term should they find it necessary.

The six most used unique terms in the answers of this test were:

- 41x none (geen)
- 8x joy (blijheid)
- 5x sympathy (medelijden)
- 5x delight (opgetogen)
- 4x fear (angstig)
- 4x resigned (berustend)
- 3x sadness (droevig)

Given the large amount of 'none' answers, it can also be concluded that a lot of samples did not contain a clear emotion. This could also explain the large amount of slightly varying terms used, but without a larger scale test (which was not the purpose of this preliminary test anyway), it is not possible to clearly confirm or deny that the amount of terms was a result of the samples not containing a clear emotion, or that it was a result of personal preferences on the terms used. Because the audio material used for this test was limited to samples of a few stories, it is likely that other stories contain emotions other than those in this list. Therefore the samples used in the preliminary tests cannot be considered a representative of the samples used in the final experiment. Also, since both 'joy' and 'delight' are tertiary emotions linked to the primary emotion 'joy', and 'sympathy' is a tertiary emotion to 'sadness', it was decided to use all the primary emotions listed in table 2.2 as possible answers in the multiple-choice test. When linking the various terms used in the preliminary experiment to their primary emotions (insofar as the translation allowed this), the emotion-count was as follows (terms that did not translate to an emotion were excluded):

- 41x none
- 27x joy
- 8x love
- 4x sadness
- 4x anger
- 4x fear
- 2x surprise

As such, these options were used in the public test: None, Love, Joy, Surprise, Anger, Sadness, Fear and Other.

4.3 Target participants

In order to get enough data to reliably base any judgements on, which are also statistically viable, the bigger the number of participants, the better. There is however the possibility of people who will be less inclined to finish the entire test and be serious and truthful about the answers even though they thought it was a good idea at the time when they started on the test. Earlier large tests involving language, like the Blizzard Challenge 2005 [1] where voices from different TTS systems were compared by various people of different backgrounds (professional linguists, students, undergraduates), showed that in order for a test to result in statistical viable results, you will need to focus on experts and people who have enough incentive to contribute positively to the test. This means that people who have no ties whatsoever to this test, neither scientifically nor socially (because they know people who are also participating in either the testing or the creation and are genuine about willing to help) will be excluded from participation in order to remove potential unhelpful (or even counter-productive) results. The professional and social aspects should be incentive enough for the participants to aid in performing this test. As a result of this, only people in the EWI department, family, and some other students were invited to participate in this test. This limits the total number of people to participate, but likely increases the overall amount of completed test runs.

4.4 Test environment

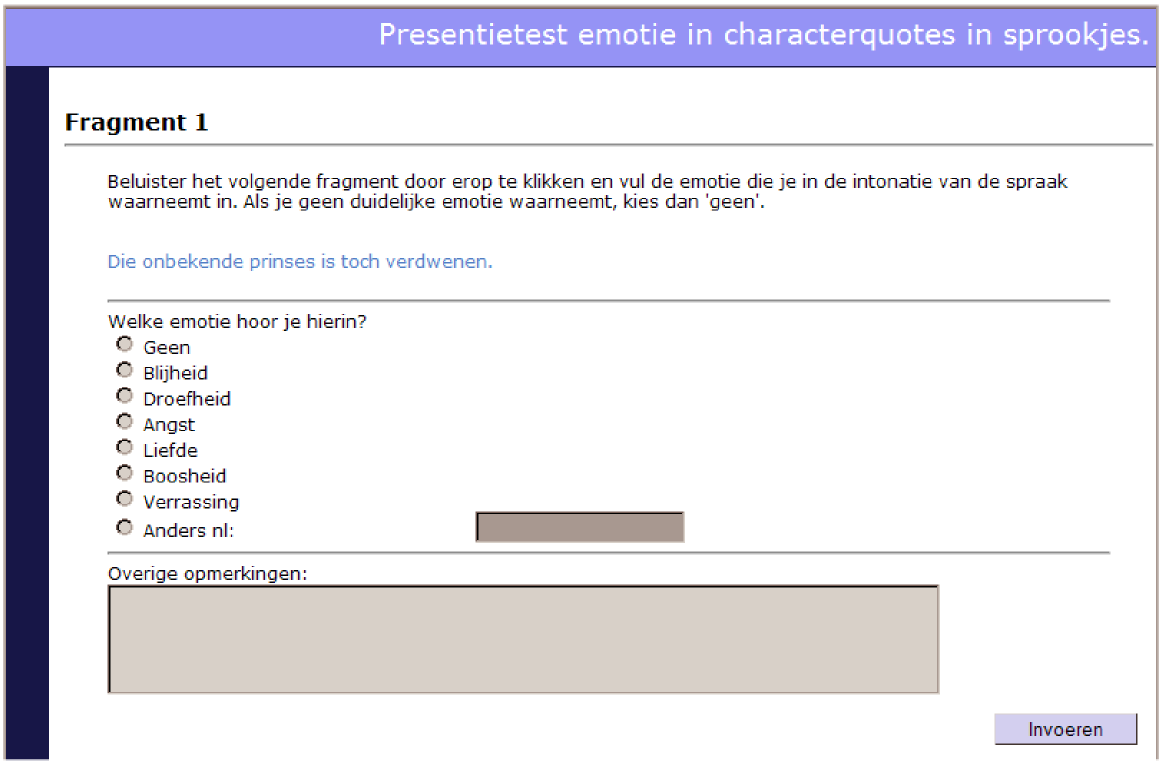
In order to reach a maximum number of participants, it is a good idea to make the test available over the Internet. This allows people who are willing to contribute to do so without having to go to the extra effort of travelling towards wherever they are needed to go in order to take the test. This did put further constraints on the test with respect to, for instance, browser compatibility. Since the test was based on audio fragments, the quality of the audio equipment people have at the machine they used could also have influenced the test to some extent. However, because the original audio came from old cassette tapes - meaning the quality was not very high to begin with - the influence of the used equipment to listen to the audio was assumed to be minimal. Measures needed to be taken for people to allow them to get used to the audio fragments used in order for them to listen to the audio fragments in an optimal way. This means that a demonstration page had to be created. This demonstration page contained a fragment which the users could use to modify the volume and other audio-related settings in order to hear the fragments in a way they thought best before actually starting with the test.

4.5 Test structure

The fragments presented to the user consisted of a piece of a story containing a character quote. The fragments were presented simultaneously as audio and text. This was done simultaneously to prevent people to think about what is being said because they could read it at the same time. This works similarly as subtitling at a tv-show or movie where people can read the subtitles without actually realising they're doing so and still focus on the show or movie itself. The user was presented with a multiple-choice answering method (which was presented at the same time as the text and audio) which contained the question ”Which emotion do you hear?” and the possible (exclusive) answers consisting of a few predetermined emotions, ”None” and ”Other”. When the ”Other” field was selected, the user was required (this was checked) to enter something in an editbox that was enabled as soon as the ”Other” field was selected.

At the bottom of the screen a ”next” button allowed the user to continue to the next test item. This button was disabled by default, and was enabled only if either ”None”, a specific emotion, or ”Other” was chosen. If ”Other” was chosen, the editbox had to be non-empty. The time it took from the page to load until the ”next” button was pressed was saved. This allowed us to measure some sort of confidence factor correlated with the answer (for example: if it takes a long time, the user is probably thinking hard on what to answer, thereby possibly decreasing the confidence of the answer itself).

A screenshot of a question of the test is shown in figure 4.1.



Presentietest emotie in characterquotes in sprookjes.

Fragment 1

Beluister het volgende fragment door erop te klikken en vul de emotie die je in de intonatie van de spraak waarneemt in. Als je geen duidelijke emotie waarneemt, kies dan 'geen'.

Die onbekende prinses is toch verdwenen.

Welke emotie hoor je hierin?

○ Geen
○ Blijheid
○ Droefheid
○ Angst
○ Liefde
○ Boosheid
○ Verrassing
○ Anders nl:

Overige opmerkingen:

Figure 4.1: Screenshot of a question of the presence test

In order to enforce the user not to spend too much time thinking on which emotion to pick (which indicates the emotion was not that clearly present), it was possible to implement a time-limit which, when a timer expired, disabled the emotion-selection input fields. This would allow the user to still enter a comment, or continue entering comments, for the specific question, but leave the user unable to choose or alter the previously made choice of the emotion perceived. It was deemed unnecessary to implement this, since the elapsed time per question was measured anyway.

The time it would take an average user to complete the test (barring excessive commenting) was aimed to be about 20 minutes. This comes down to about 20 fragments with a length varying from as little as 1 up to 10 seconds, which would leave the user about 55 seconds on average to supply an answer. The fragments were supplied in the same fixed order for all users. This was to insure that at least the starting fragments would have a statistical significance in the possible event that users did not finish the test. For statistical purposes, it is better to have a small number of fragments labelled by many, than a large number of fragments labelled by few. In order to avoid fragments originating from the same story ending up right after each other (and thereby giving the user too much context information about the underlying story which could lead to easy emotion prediction) the order was shuffled once at the initialisation of the test server and checked afterwards.

4.6 Experiment results

The test was performed by 32 people, who all completed the entirety of the test. All subjects were native Dutch speakers. The fragments used in the test were collected from 3 fairy tales: "Assepoester", "Dot en de kangeroe", and "De drie wensen". The sentences, the name of the character that speaks them, the internal sample name (which contains the name of the fairy tale the fragment originated from) are shown in table A.1. The names of the fairy tales have been abbreviated with: a, dedk and ddw (the first letter of every word).

Table A.2 contains the frequency of the various answers per question in the order in which each fragment appears in the story. As can be seen at first glance, there usually is one emotional term that has been chosen more than the others. Also, the amount of 'other' answers is reasonably high for the characters Magda, Frederik and Tante. Table A.3 lists the full fragmental text, as well as the list of 'other' results given (in Dutch).

It was surprising to see that, despite built-in checks that would not allow someone to continue to the next question if 'other' had been selected while no text had been entered in the accompanying textbox, some people managed to do just this. Also, some people entered quite a lot of text in the 'other' emotion textbox instead of the comment field which resulted in loss of text due to database design constraints. The 'other' emotion textbox was limited to 30 characters (which should be enough to contain any single emotional term), while the comment area was unrestricted in size. This happened 4 times, all by the same person. He or she phrased the 'other' emotions as choices between two emotions and continued to provide contingency answers. For example: "Enthousiasme, anders verrassing" or "Opgelucht; als dat geen emotie" (note the 30 character limit). In such a case, the first mentioned emotion was used as the answer.

In an attempt to emphasise the statistical weight of various answers, it has been decided to count emotions listed under 'other' towards the primary emotion listed in the emotions table 2.2. This is done under the assumption that if emotions are closely related, the method of expressing and intensity of these emotions is probably closely related as well. This resulted in 33 'other' answers being moved to one of the listed primary emotions. There were also a couple of answers that were moved to another emotion. These are shown in table 4.2. The phrase "wat nou weer?" is a clear indication of irritation, which is a secondary emotion to anger. "geruststelling" translates to appeasing which, in my opinion, is an expression of compassion, which is a tertiary emotion to love.

The last two entries where emotions were changed were answered in the presence test in such a way that either emotion could be applied. I moved these answers to the category that already had the most answers. For the 'Blijheid-Verrassing' change, there were 3 other people who answered 'Blijheid' and 15 others who answered 'Verrassing'. For the 'Liefde-Droefheid' change, there were 4 other people who answered 'Liefde' and 18 who answered 'Droefheid'. These figures were enough

Old emotion	**New emotion**	**Comment**
Surprise	Anger	vraag toon, op toon van " wat nou weer?"
None	Love	soort geruststelling, maar ik kan het niet echt koppelen aan een emotie
Happiness	Surprise	Hangt een beetje tussen blijheid en verrassing in
Love	Sadness	Ook droefheid

Table 4.2: remaining emotion modifications

to convince me that these 2 in-between answers could safely be merged into the larger category.

Modifying the emotions in the above 4 examples and the 33 'other' answers led to slight changes in the agreement values per question, but did not change the overall outcome. Table A.4 lists the question number, the agreed-upon emotion, and the agreement factors (as a value between 0 and 1) before and after moving some of the 'other' emotions to their primary emotion. Before the modification, 132 'Other' answers were present, afterwards, 99 remained. Table A.5 contains the list of comments and what emotion they were changed to. Here is a count of which emotions the 'other' descriptions were moved to:

- 19x Anger
- 4x Love
- 5x Surprise
- 3x Joy
- 1x Sadness
- 1x Fear

In order to find out whether the test participants confused specific emotions, and if so, which emotions, and to what degree, a coincidence matrix was created. The figures shown in table 4.3 are percentual. The values show the confusion, averaged over all questions. The higher the figure, the more people disagreed between the two emotions (row and column), except for the diagonal elements. There the figure is a representation of how much people agreed on the emotion. This shows that the most confusion comes from some people claiming a fragment had no emotion, and other people claiming it had (totalling to 22.08% of all the cases where, for a single fragment, someone picked "None" and someone else picked an emotion). When looking at specific (so not counting 'Other') emotions being confused, the percentage totals to 13.22%. Of this percentage, the main contributors are the pairs: Fear - Sadness (2.87%), Surprise - Anger (2.15%), and Anger - Sadness (2.11%). On the whole, the main source of confusion comes from the 'None' vs any emotion (22.08%), and from 'Other' vs any emotion (including 'None') with 12.08%. The confusion between 'None' and 'Other' was 4.54%. This leads to the suspicion that the fragments did not contain as many prominent emotions as was initially expected, or the emotions were not that prominent that they stood out enough.

	None	**Happiness**	**Sadness**	**Fear**	**Love**	**Anger**	**Surprise**	**Other**
None	14.86	3.08	2.84	2.92	2.02	3.13	3.54	4.54
Happiness	3.08	6.04	0.70	0.37	0.12	0.17	1.23	1.30
Sadness	2.84	0.70	9.28	2.87	0.88	2.11	0.16	1.64
Fear	2.92	0.37	2.87	5.57	0.58	0.36	0.88	1.26
Love	2.02	0.12	0.88	0.58	0.78	0.42	0.23	0.36
Anger	3.13	0.17	2.11	0.36	0.42	11.51	2.15	1.84
Surprise	3.54	1.23	0.16	0.88	0.23	2.15	6.42	1.86
Other	4.54	1.30	1.64	1.26	0.36	1.84	1.86	2.01

Table 4.3: Relative coincidence matrix of given answers in % (N=32, Items=30)

In order to determine the degree of agreement amongst the participants, Krippendorff's [25] alpha-value was used. This method uses a statistical analysis that results in an alpha value. The value of alpha is a measure of reliability of the results of the test. A value of 1.0 indicates perfect reliability, whereas a value of 0.0 indicates that the participants made up their answers by throwing dice. The alpha-value of the presence experiment turned out to be 0.284. As such, this indicates that the test participants are in quite poor agreement on the emotion perceived in the fragments.

Another way to look at this is to calculate the average agreement per emotion. The percentages were averaged per emotion, the results are shown in table 4.4. The percentages for all questions are also shown in figure 4.2. The horizontal axis shows the fragments, and for each dataset the emotion that was perceived the most, ordered from best-match (left) to worst-match (right).

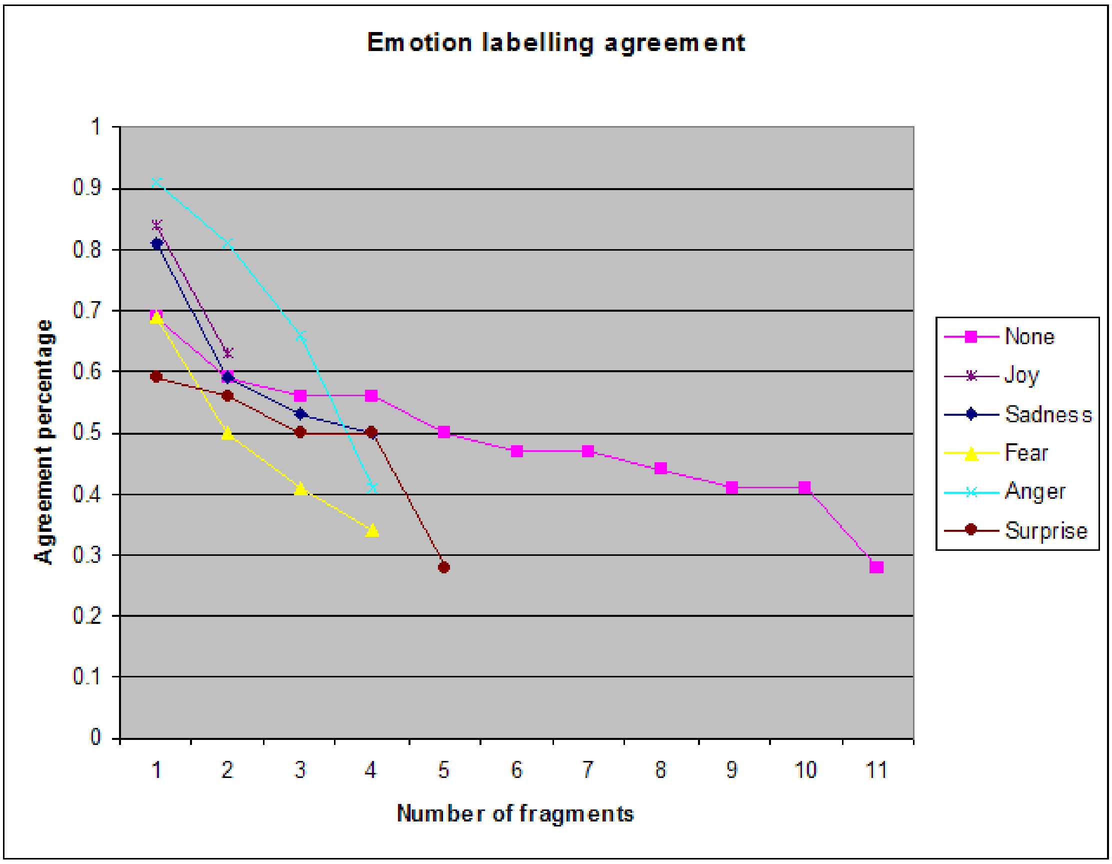


Figure 4.2: Agreement on emotion per fragment

Emotion	None	Joy	Sadness	Fear	Anger	Surprise
Average Percentage	48.9%	73.5%	60.8%	48.5%	69.8%	48.6%

Table 4.4: Average agreement percentages per emotion

As can be seen in the figure, from the range of emotions only a few fragments are agreed upon by less than 50%. Only "no emotion" (None) has most fragments rated with an agreement below 50%. But 50% agreement on one option is, given the choice of eight options, well above chance level (12.5%)

Putting the emotions from the test side by side with the emotions determined from the context of each fragment results in table 4.5 (the fragments are ordered in chronological order per fairy-tale):

It is interesting to see that when comparing the emotions resulting from the test matched with the emotions ascertained from the context in 18 of the 30 fragments (60%). Also, in 2 fragments,

Character	**Sentence**	**Test**	**Context**
Tante	Wat jij moet aantrekken?	Anger	Anger
Tante	Dat gebeurt misschien nog wel	None	None
Tante	Die onbekende prinses is toch verdwenen	None	None
Tante	Ik zal het eerst proberen	None	None
Frederik	Wat is er met jou aan de hand?	Surprise	Anger
Magda	Kom binnen en doe de deur dicht	Fear	Fear
Frederik	Ik zou best een paar worstjes lusten	None	None
Magda	Dat wens ik ook	None	None
Magda	Er is een brief van de feeen gekomen	Surprise	Happiness
Magda	En daarin staat dat we drie wensen mogen doen	Happiness	Happiness
Frederik	We moeten heel goed nadenken Magda	None	None
Magda	Ik heb al een lijstje gemaakt	None	None
Magda	Luister wat ik heb bedacht	None	None
Frederik	Wat? Is er geen eten	Surprise	Anger
Frederik	Wat ben je toch lui Magda	Anger	Anger
Magda	Wat ben je toch dom	Anger	Anger
Frederik	Waarom hebben we ook zo'n ruzie gemaakt over die wensen	Sadness	Sadness
Frederik	Nee, nee, het was jouw schuld niet lieverd	None	Love
Dot	Ja ik ben de weg kwijtgeraakt.	None	Sadness
Dot	Maar dat is de baby van mijn kangeroe	Anger	Surprise
Dot	Ik vind het anders helemaal niet grappig	Sadness	Anger
Dot	Ik heb nog nooit een vogelbekdier gezien	Surprise	None
Dot	Maar iemand moet toch weten waar mijn huis is?	Fear	Fear
Dot	Wie zijn dat?	Fear	None
Dot	Ik heb mensen nog nooit zo raar zien doen	Surprise	Fear
Dot	Kangeroe, laat mij maar uitstappen	Fear	None
Dot	Laat mij toch hier	Sadness	Sadness
Dot	Oh lieve kangeroe	Sadness	Sadness
Dot	Dank u wel voor uw hulp	Happiness	Happiness
Dot	Ik u ook	None	Sadness

Table 4.5: Comparing emotions from the test with the context of the fragments

of which the users agreed to a specific emotion with 81.25% agreement, the emotion perceived did not match the emotion ascertained from the context. This was the case with the sentences "Maar dat is de baby van mijn kangeroe", where anger was perceived but the context revealed surprise, and "Ik vind het anders helemaal niet grappig" where sadness was perceived but the context showed that statement was made while being angry.

The fragments to be used in the detailed acoustic analysis were filtered on the following requirements:

- The level of agreement on a specific emotion of the fragment must exceed random chance (12.5%), the higher, the better.
- At least one of the fragments per character must be emotionally neutral.

The first requirement discounted none of the fragments, and the second requirement also did not discount any fragments. Concluding, because all the fragments met the above criteria, they were all used for analysis with which a neutral-to-emotion conversion model was constructed. This resulted in 2-4 fragments per emotion over all characters. However, the poor value of Krippendorff's alpha implicated that the emotions were not present that clearly; at least not clear enough for the majority of the test users to agree upon a single emotion. This means that the data resulting from an acoustic analysis could result in equally non-agreeing results.

4.7 Discussion

An alternative approach to obtaining fragments of emotional speech of which the emotion is known, is by employing a (voice) actor. A professional actor is able to simulate emotions in his/her speech to such a degree that it is mostly indistinguishable from 'normal' spontaneous emotional speech. Using an actor as a source of emotional speech bypasses the problem of having to label the recorded speech fragments with the emotions present; those were already known in advance. However, this deviates from the plan to use speech from commonly accepted storytellers. Another method of obtaining emotional speech is by using a corpus of spontaneous emotional speech. The downside to this approach is that the corpus consists of speech from a large amount of different people, making it harder to find both neutral and emotional speech fragments from the same speaker. This also deviates from the plan to use speech from commonly accepted storytellers, because a corpus usually consists of speech fragments from a diverse range of people.

4.8 Conclusion

The experiment setup proved adequate. All people who participated finished the test. The agreement percentages per emotion in figure 4.2 were all well above chance level and therefore statistically sufficient to label each fragment with the emotion that was perceived by most of the experiment participants. The Krippendorff alpha value of 0.284 does, however, give an indication that people did not agree much with each other on a global scale. This is probably caused by the fact that emotions are extremely subjective, and people are not always equally good at recognising and formulating the perceived emotions as well as agreeing on the perceived emotions. What sounds surprised for one apparently sounds angry for another. These conflicting perceptions are a major contributor to the low alpha value.

Chapter 5

The birth of a model

In this chapter, the results of the acoustic analysis are presented. First, specific aspects of speech which have been used to analyse speech in other similar research available in literature are discussed in section 5.1. Then, the process of measuring and analysing the data is described in section 5.2. The results are then compared to the ones found in literature in section 5.3 and conclusions are made in section 5.5. The result is a model that, given an emotion, will return the necessary acoustic modifications needed in order to transform a neutral voice into one that carries the given emotion.

5.1 Speech analysis in literature

In earlier research done by Jianhua Tao et al. [53], Eva Navas et al.[35], Rank & Pirker [41] and Sylvie Mozziconacci [31] several interesting prosodic attributes for analysing emotion in speech utterances are used. The research in these works contains detailed *absolute* measurements, which promised to be useful when the need to comparing measured prosodic attributes arised. More on comparing measurements can be read in section 5.2. Since the first three mentioned research projects focused on the Chinese, Basque and German languages respectively, not all the employed attributes are usable in an analysis of the Dutch language as well, but there are certainly some attributes that do:

- Syllable duration
- Maximum value of the pitch curve and its first derivative
- Mean value of the pitch curve and its first derivative
- Range (Max-Min) of the pitch curve and its first derivative
- Standard deviation of the pitch curve and its first derivative
- Maximum value of the intensity levels and its first derivative
- Mean value of the intensity levels and its first derivative
- Range of the intensity levels and its first derivative
- Standard deviation of the intensity levels and its first derivative
- The amount of F0-jitter
- The amount of intensity-shimmer
- The harmonics-to-noise ratio
- The F0 value at the end of a sentence (EndFrequency)

Here, F0-jitter is used to describe the variation of F0 from one pitch period to another. Similarly, shimmer is related to microvariations in the intensity-curve. The harmonics-to-noise ratio, also known as voice creakiness, is related to spectrum of the glottal excitation signal. Longer pulses have less high-frequency energy than shorter pulses. A creaky voice is the result of skipped (or missing) pulses. Unfortunately, Navas and Tao measure jitter and shimmer in different ways. Navas uses the amount of zero-crossings of the derivative curves, while Tao measures jitter as follows:

> For F0 jitter, normally, a quadratic curve is fitted to the acoustic measurement with a moving window covering five successive F0 values. The curve is then subtracted from the acoustic measurements. F0 Jitter was calculated as the mean pitch period-to-period variation in the residual F0 values.

Praat uses yet another way of measuring jitter and shimmer, and so the different measurements cannot be compared to one another.

5.2 Speech analysis

Firstly, all the audio fragments made available for the presence test were converted from their original source. The sample format used in the presence test was: Microsoft ADPCM 4bit, 22050Hz, mono. These samples were also normalised on their volume to avoid having some samples being significantly louder (or softer) than others, which is a source of great annoyance (and distraction) when having to listen to a large number of samples. The measurements, however, were done using the original samples (PCM signed 16bit, 22050Hz, stereo). All the stories were the samples are from were recorded into one sample per story. This means that all the samples that come from a single story are all recorded at the same volume levels.

The fragments were annotated with the words spoken, and data measurements were taken using Praat [39]. Praat does not handle stereo samples, and takes only the first channel available. Fortunately, the channels were almost completely the same, so measurements are not affected by this. Measurements consisted of F0-values and intensity-values of the audio signal per word. Pauses, and other non-speech sounds like drawing breath were not measured. Pauses can, to some degree, be derived from the period of vocal inactivity between words. Non-speech sounds cannot be recreated by the TTS system and were thus not measured. The measurements were saved as series of time-value pairs, where F0 was measured in Hz and intensity in dB. Jitter, shimmer and the harmonics-to-noise ratio were measured using all the data from the start of the first word until the end of the last word per sentence.

Praat shows the amount of jitter in 5 different ways but because no other data was found in publications that use the same method of measuring jitter as Praat does, the choice was made to include jitter (local) since that was the default setting[1] in the measurement data. The same applies to shimmer[2].

The values of the time-axis in the measured time-value pairs for F0 and intensity produced by Praat's analysis were asynchronous. In order to synchronise this, the data was imported into matlab, where the data was interpolated using matlab's interp1-function over the union of the 2 asynchronous sets of time values.

The mentioned analysis items, concerning F0 and intensity levels, were then measured on the interpolated data per word in order to be able to analyse if the acoustic attributes of specific words aided more significantly in the perception of an emotion or not. In order to measure the amount of F0-accent put on each word, the average and max F0 values of each word were compared with the overall F0 trend line of the entire sentence. In order to obtain this trend line, a method was constructed to calculate this trend line based on the measured F0 contour. The aim was to find an algorithm that uses the F0 measurements as input, and produces a set of F0 measurements that match the declining F0 trend line as much as possible. This was done, for regular declarative sentences, by calculating a first-order regression line of all F0 values of the sentence. Because this regression line includes the values of the accented words, the line is actually above the trend line. This can be seen in Figure 5.1 where the F0 contour is given in red, the regression line is shown in blue, and the trend line is green. By creating a new regression line which uses all the values of the original F0 contour *except* those values that are higher than the regression lines, the trend line

[1] The Praat help file states: "This is the average absolute difference between consecutive periods, divided by the average period. MDVP calls this parameter Jitt, and gives 1.040% as a threshold for pathology."

[2] The Praat help file states: "This is the average absolute difference between the amplitudes of consecutive periods, divided by the average amplitude. MDVP calls this parameter Shim, and gives 3.810% as a threshold for pathology."

is created. This method does not work for sentences that use a non-falling F0 contour, or those that have a falling and then a rising F0 contour. It is theorised that calculating the trend line for each F0 segment (falling, rising etc) could result in better trend lines, but this was not further pursued. The sentence in the figure is "Er is een brief van de feeën gekomen". By comparing the mean F0 with the mean trend line values of each word, we get the following values:
Er (101%) is (98%) een (98%) brief (132%) van (107%) de (86%) feeën (158%) gekomen (100%).
This nicely shows the accentuated words "brief" and "feeën", as well as the 'muffled' pronunciation of "de". The maximum values are (obviously) a little higher all around:
Er (107%) is (101%) een (104%) brief (161%) van (111%) de (93%) feeën (176%) gekomen (118%).

If this were analysed in conjunction with a pos-tagger and/or other grammar related rules, it is my opinion that this could result in useful information about how certain emotions are realised through the manipulation of specific words. This does not restrict to just the F0 contour and trendline, but for example also to word duration, pause duration and intensity levels. However, investigating this was expected to take too much time and focus away from the main research, and was left as future work.

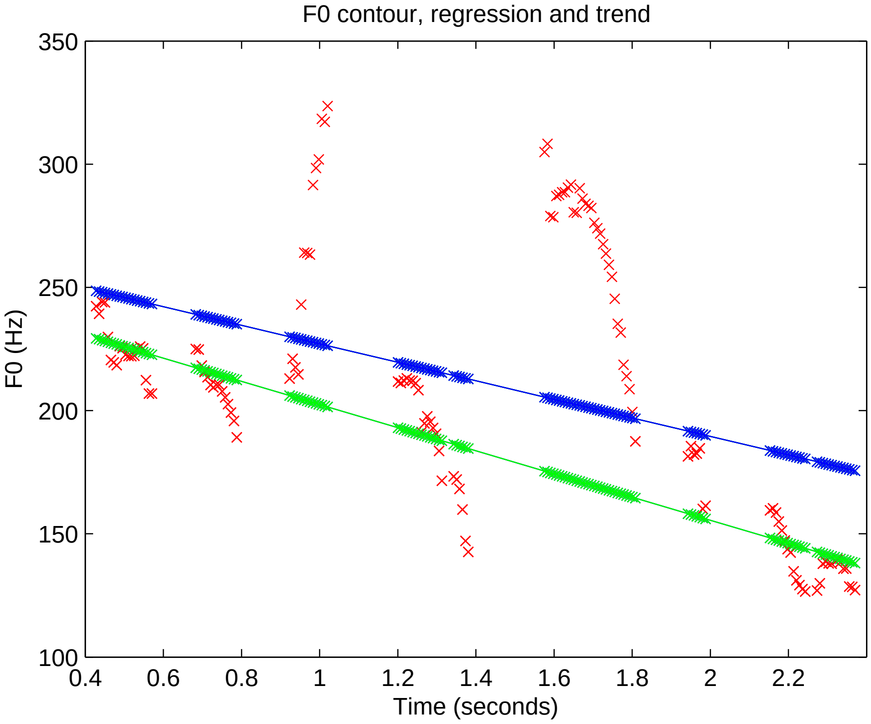


Figure 5.1: F0 contour (red), regression line (blue) and trend line (green) of the sentence "Er is een brief van de feeën gekomen"("A letter from the fairies arrived"

In order to extract per-sentence measurements, the unmodified per-word data elements were concatenated to form sentences (note that inter-word pauses were not analysed, but the time-axis is consistent per fragment, and thus per sentence). These values were then interpolated like the per-word data but with one notable difference. The time-axis the data was interpolated to was not a union of both the time values of the F0 measurements and the intensity measurements, but a set of fixed intervals over the same time period as the F0 and intensity measurements. This was done in order to get more accurate measurements for the derivatives of the F0 and intensity values. When calculating the derivative dy/dx, less erratic results are obtained if dx remains constant. Speech rate was calculated by subtracting the first time value of the first word from the last time value of the last word and dividing the result by the number of syllables in the sentence.

Tables 5.1 to 5.5 show the results of the measurements performed on the fragments per fragment, per character. The emotion labelling is the one ascertained by the presence test. In order to show to what degree the values differ (or not) per emotion for each measured element, figures were created. In these figures, for example Figure 5.2, each measured element is shown on the x-axis, where the various emotions are shown on the y-axis. As can be seen, the emotions 'None' and 'Surprise' (verrassing) have a similar mean F0 range, while 'fear' (angst) has a range so large that it encompasses the remaining values for the other emotions. Figure 5.3 shows the measurements of the F0 range per emotion. The collection of figures for each character and attribute is available in the digital appendix.

Sentence	20	17	27	2	24	21
Emotion	Fear	Fear	Fear	Sadness	Sadness	Sadness
Max F0 (Hz)	491	433	494	477	415	469
Mean F0 (Hz)	348	331	402	381	356	387
F0 Range (Hz)	220	192	202	202	160	155
StdDevF0 (Hz)	61	59	55	64	38	38
Max dF0 (Hz/sec)	4417	3182	2788	5263	1184	3342
Mean dF0 (Hz/sec)	11	-119	-79	-170	-340	-70
dF0 Range (Hz/sec)	8317	7081	6688	9163	5084	7242
StdDevdF0 (Hz/sec)	997	1250	1241	1107	772	917
Max Int (dB)	84.27	78.7	83.77	83.07	82.93	84.77
Mean Int (dB)	77.66	73.45	75.34	75.98	76.85	74.86
Int Range (dB)	21.84	18.86	24.25	24.74	23.26	22.75
StdDev Int (dB/sec)	4.13	4.09	5.84	5.18	4.68	5.2
Max dInt (dB/sec)	656.44	499.73	865.05	561.92	572.29	550.22
Mean dInt (dB/sec)	-14.41	-25.78	-41	-48.94	-24.32	-24.18
dInt Range (dB/sec)	747.97	588.45	1019.45	960.45	791.94	579.42
StdDev dInt (dB/sec)	176.71	199.33	252.25	239	207.62	163.63
Jitter (local)	1.96%	2.85%	3.73%	2.01%	2.22%	3.06%
Shimmer (local)	7.93%	13.08%	12.30%	9.14%	8.46%	11.98%
Fraction of locally unvoiced frames	19.33%	27.27%	11.32%	21.86%	12.84%	3.14%
Mean harmonics-to-noise ratio (dB)	12.917	6.004	6.663	11.791	10.67	9.44
Speech Rate (Syll/sec)	5.86	4.81	5.35	6.93	4.55	4.6
End Frequency (Hz)	455	315	298	287	349	383

Table 5.1: Analysis of fragments by character Dot part 1

For the character Dot (with the most fragments in total), the emotions anger and sadness each were linked to three fragments, and the emotions surprise and 'none' were each linked to two fragments, leaving the emotions happiness and anger with only one fragment each. This number of fragments, is too few to create a good many-to-many relation[3] between a neutral sounding fragment ('none' emotion) and any other emotion. Also, a lot of the data measured is quite inconsistent within the different characters. as can be seen reflected by the large spread of the values for, for example, the emotion fear ('angst') in figure 5.2. If these values were clustered closely together, then it would be possible to correlate that specific value range to an emotion, but with ranges this wide, the current amount of fragments is not enough to demonstrate a correlation. This applies to all the attributes measured for all characters, the two figures 5.2 and 5.3 are the most insightful because the character Dot had the most fragments. The character Tante had only four fragments, and only one of those was labelled an emotion other than 'None', leaving little to

[3] linking two sets of values to each other

Sentence	8	18	6	16	14	23
Emotion	Happiness	Anger	None	None	Surprise	Surprise
Max F0 (Hz)	519	548	414	344	528	410
Mean F0 (Hz)	351	389	305	296	281	312
F0 Range (Hz)	301	297	218	117	358	187
StdDevF0 (Hz)	81	75	63	38	107	54
Max dF0 (Hz/sec)	5509	6328	2642	1442	7961	2302
Mean dF0 (Hz/sec)	63	-181	-321	-401	168	-142
dF0 Range (Hz/sec)	9409	10228	6541	5342	11861	6202
StdDevdF0 (Hz/sec)	1463	1044	1037	645	1318	803
Max Int (dB)	83.53	84.36	82.37	78.48	82.56	80.32
Mean Int (dB)	77.51	77.33	75.3	73.34	73.52	72.17
Int Range (dB)	20.38	26.43	21.84	16.46	24.81	22.89
StdDev Int (dB/sec)	4.8	5.41	4.72	4.4	4.63	3.85
Max dInt (dB/sec)	520.48	719	481.16	335.91	681.32	698.13
Mean dInt (dB/sec)	-42.63	-59.41	-45.71	-67.54	-38.59	-33.76
dInt Range (dB/sec)	828.05	891.6	674.18	455.58	1010.52	951.04
StdDev dInt (dB/sec)	181.03	221.09	193.21	115.93	205.06	195.99
Jitter (local)	2.06%	1.72%	2.89%	2.80%	2.48%	2.53%
Shimmer (local)	8.09%	7.59%	13.11%	12.31%	10.64%	11.42%
Fraction of locally unvoiced frames	8.02%	14.38%	32.56%	26.36%	20.66%	19.75%
Mean harmonics-to-noise ratio (dB)	11.271	14.556	7.236	9.192	9.306	9.93
Speech Rate (Syll/sec)	5.02	5.39	4.07	4.87	5.27	4.38
End Frequency (Hz)	397	287	280	228	492	244

Table 5.2: Analysis of fragments by character Dot part 2

be shown in a figure like 5.2. All the figures are, however, available in the digital appendix which is located on the cd that comes with this report.

The wide ranges of the prosodic attributes measured can be caused by the fact that the emotions are really expressed differently (or were not the same emotion to begin with), the labelling could be wrong, or the measured attributes simply don't contribute to some specific emotions in an easily discernible way. Despite these incoherent measurements, it is possible to extract a linear relation between each emotion and the 'None' emotion per character. This is done by dividing the average values of a category linked to an emotion by the average values of the same category linked to the 'None' emotion. The results of this can be found in tables A.7to A.10 where all the absolute values have been replaced by relative values with respect to the average values of the 'None' emotion. The 'None' emotion has been left out of these tables, as the values would obviously result in only ones. A short extract of these tables which shows only a few of the measured prosodic attributes of the utterances can be seen in Table 5.6. The numbers here are simple multiplication factors to be used in the formula: $Attribute(target_emotion) = Attribute(neutral) * Factor(target_emotion)$ The numbers of the other measured prosodic attributes of all the characters were calculated the same way. These numbers do not show anything significant on their own, but will be used in the next section in the comparison with measurements from literature.

Averaging the values like this will result in a crude model, but only a large amount of additional analysable data could fix this.

Of all these different measured categories, only a few are directly manipulable in festival: intonation, intensity and duration. This means that we will only be able to use the measurements on F0, intensity and speech rate.

Sentence	28	4	30	22	13	29	11
Emotion	Fear	Happiness	Anger	None	None	None	Surprise
Max F0 (Hz)	318	406	482	303	411	376	322
Mean F0 (Hz)	254	234	402	224	249	245	213
F0 Range (Hz)	125	281	221	156	252	207	195
StdDevF0 (Hz)	28	87	66	46	75	51	52
Max dF0 (Hz/sec)	3197	5221	2998	3537	3266	1360	3773
Mean dF0 (Hz/sec)	-84	-95	-181	72	104	-526	-216
dF0 Range (Hz/sec)	7097	9121	6897	7437	7166	5260	7673
StdDevdF0 (Hz/sec)	964	1041	1211	705	1164	845	898
Max Int (dB)	75.18	79.12	83.03	81.08	80.46	74.81	73.17
Mean Int (dB)	69.70	69.08	73.99	70.98	70.60	68.88	65.76
Int Range (dB)	17.68	21.92	26.91	26.94	23.54	19.33	17.06
StdDev Int (dB/sec)	4.11	5.55	6.74	5.73	6.47	4.68	2.90
Max dInt (dB/sec)	234.25	629.95	643.33	485.13	692.21	667.31	428.63
Mean dInt (dB/sec)	-71.25	-61.40	-81.75	-63.36	-85.10	-100.81	-55.85
dInt Range (dB/sec)	579.05	843.03	839.88	921.05	977.00	941.32	665.02
StdDev dInt (dB/sec)	150.79	204.95	246.24	242.48	317.15	222.24	166.76
Jitter (local)	3.35%	3.01%	1.92%	3.45%	4.29%	3.78%	3.51%
Shimmer (local)	13.47%	11.51%	9.74%	12.66%	13.79%	14.82%	14.44%
Fraction of locally unvoiced frames	35.19%	27.62%	21.52%	24.51%	35.75%	33.77%	25.00%
Mean harmonics-to-noise ratio (dB)	8.983	9.546	12.416	7.951	5.834	7.624	6.785
Speech Rate (Syll/sec)	4.48	4.33	5.25	6.43	6.48	4.35	5.53
End Frequency (Hz)	245	127	431	149	160	171	127

Table 5.3: Analysis of fragments by character Magda

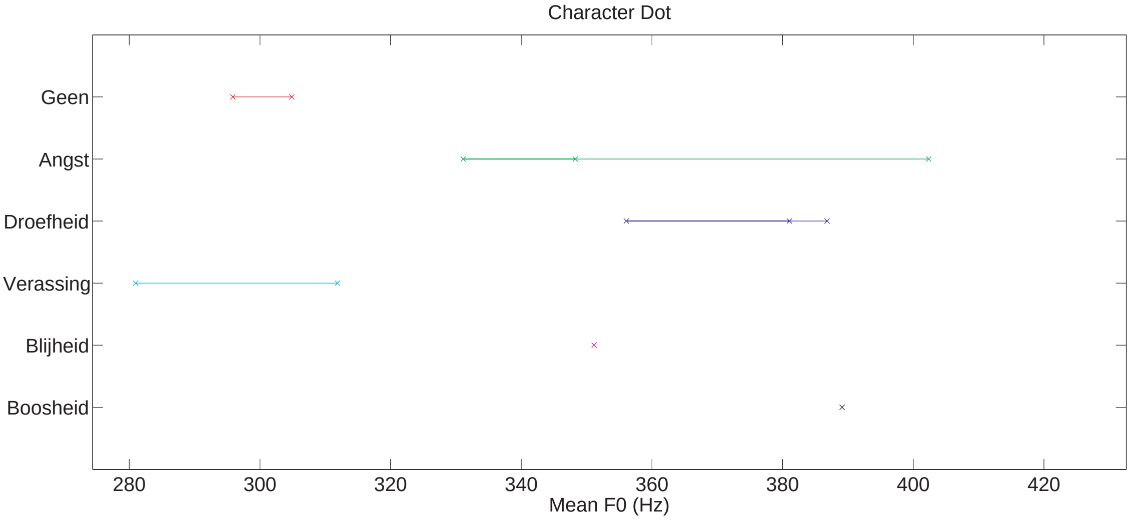


Figure 5.2: Mean F0 values of sentences by character Dot

5.3 Comparison with other measurements

As mentioned earlier in section 5.1, the available literature also provides a source of earlier measurements to compare the measurements obtained in section 5.2 with. Unfortunately, not everybody uses absolute measurements, and in order to find more literature to compare the measurements with, the search was extended to include literature with relative measurements. Marc Schröder

Sentence	7	10	19	9	25	5	12
Emotion	Anger	Sadness	None	None	None	Surprise	Surprise
Max F0 (Hz)	184	320	210	153	251	244	351
Mean F0 (Hz)	147	231	154	135	166	185	214
F0 Range (Hz)	69	141	95	40	115	116	245
StdDevF0 (Hz)	17	24	30	9	36	40	84
Max dF0 (Hz/sec)	1896	3971	1782	792	861	940	2178
Mean dF0 (Hz/sec)	-90	-59	-5	-9	-104	30	143
dF0 Range (Hz/sec)	5795	7871	5682	4692	4760	4840	6078
StdDevdF0 (Hz/sec)	359	811	393	216	412	287	879
Max Int (dB)	81.01	82.23	81.53	77.42	79.49	82.39	83.65
Mean Int (dB)	72.18	71.81	71.58	69.69	68.14	76.64	69.45
Int Range (dB)	22.13	23.99	24.93	22.31	21.70	25.46	27.83
StdDev Int (dB/sec)	4.70	4.68	5.87	3.60	5.77	4.89	8.54
Max dInt (dB/sec)	599.70	685.39	856.91	386.55	267.01	605.08	185.47
Mean dInt (dB/sec)	-29.11	-7.80	-27.52	-27.75	-51.92	-34.58	-94.83
dInt Range (dB/sec)	554.74	731.12	817.40	866.45	557.29	852.40	765.83
StdDev dInt (dB/sec)	169.80	212.52	175.87	148.36	165.36	188.13	185.96
Jitter (local)	2.71%	2.54%	2.10%	1.47%	4.15%	1.36%	4.81%
Shimmer (local)	12.26%	10.94%	7.40%	9.15%	18.52%	9.36%	14.13%
Fraction of locally unvoiced frames	21.43%	8.19%	17.23%	19.62%	39.39%	10.59%	39.56%
Mean harmonics-to-noise ratio (dB)	7.267	11.05	11.17	10.323	3.511	12.033	4.317
Speech Rate (Syll/sec)	4.69	5.88	4.28	3.60	4.48	7.14	3.36
End Frequency (Hz)	116	221	184	129	143	227	115

Table 5.4: Analysis of fragments by character Frederik

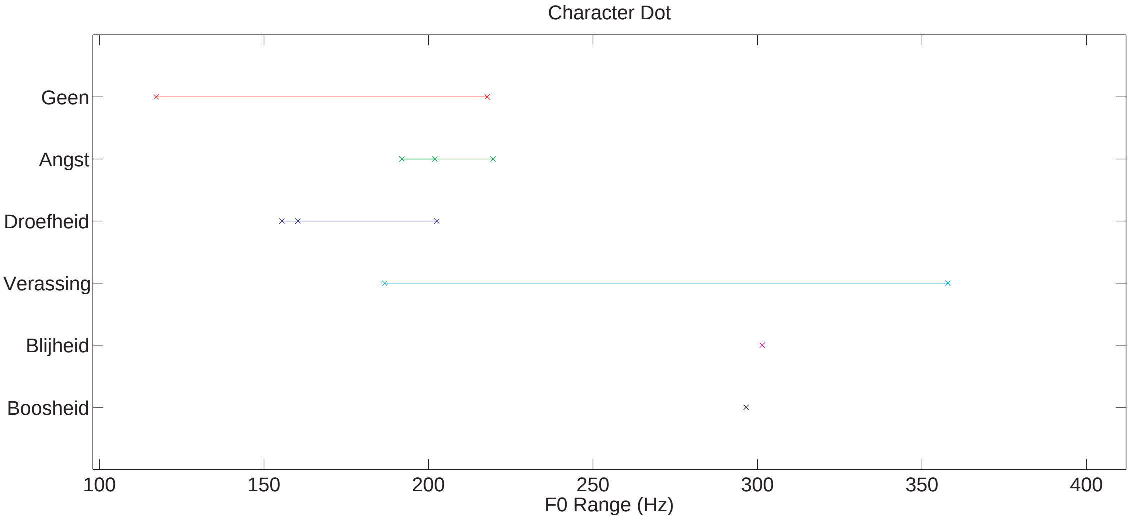


Figure 5.3: F0 ranges of sentences by character Dot

[46] has performed research in the field of emotional speech with regard to TTS systems. In his PhD thesis, which contains a large amount of literature research, he includes 11 lists of prosody rules for use in emotional speech synthesis systems presented in previous publications. This list has been copied below. Additionally, research done by Jianhua Tao et al. - who also investigate the acoustic attributes of the emotions fear, anger, sadness and joy - has been added to create the following list of literature which contain measurements of emotional speech which can be (and in

Sentence	15	3	1	26
Emotion	Anger	None	None	None
Max F0 (Hz)	484	436	191	501
Mean F0 (Hz)	274	173	157	295
F0 Range (Hz)	349	316	63	388
StdDevF0 (Hz)	114	51	16	168
Max dF0 (Hz/sec)	4130	1459	2968	2530
Mean dF0 (Hz/sec)	3	-187	-76	-293
dF0 Range (Hz/sec)	8030	5359	6868	6430
StdDevdF0 (Hz/sec)	1601	545	593	897
Max Int (dB)	80.31	78.06	76.03	78.56
Mean Int (dB)	73.72	69.87	69.95	67.86
Int Range (dB)	26.40	23.39	17.52	22.30
StdDev Int (dB/sec)	5.13	4.88	3.25	6.72
Max dInt (dB/sec)	591.59	602.11	519.41	810.14
Mean dInt (dB/sec)	-54.25	-36.72	-28.16	-46.01
dInt Range (dB/sec)	738.17	956.91	764.43	683.27
StdDev dInt (dB/sec)	229.14	221.84	174.82	255.53
Jitter (local)	3.33%	3.31%	2.80%	3.18%
Shimmer (local)	11.16%	12.76%	13.39%	7.17%
Fraction of locally unvoiced frames	22.54%	32.67%	23.53%	41.83%
Mean harmonics-to-noise ratio (dB)	7.086	1.196	9.06	7.785
Speech Rate (Syll/sec)	6.43	4.69	6.25	7.68
End Frequency (Hz)	356	151	129	114

Table 5.5: Analysis of fragments by character Tante

most cases, is) used to create prosody rules for emotional TTS systems[4]:

- Burkhardt & Sendlmeier [4] for German
- Cahn [5] for American English
- Gobl & Ni Chasaide [20] for Irish English
- Heuft et al. [21] for German
- Iriondo et al. [22] for Castillian Spanish
- Campbell & Marumoto [7] for Japanese
- Montero et al. [28] [29] for Spanish
- Mozziconacci [30]; Mozziconacci & Hermes [32] for Dutch
- Murray & Arnott [34] for British English
- Murray et al. [33] for British English
- Rank & Pirker [41] and Rank [40] for Austrian German
- Tao et al. [53] for Mandarin

The emotions appearing in the publications of the above list are mostly the same as the six emotions investigated in this thesis. The only exceptions are the emotions "Boredom" which appears only in some of the publications and was not present as an option in this thesis, and "Love" which was not in any of the publications but is present in this research. Another point is that a lot of different scales and units are used in the above list of publications. This makes comparing them a rather imprecise process. Nevertheless, the prosody rules from the publications in the list above have been copied from Schröder's work (appendix A in [46]) into the tables in appendix A.11 through A.25, with the addition of the results of Tao et al. and the exclusion of the results of Gobl & Ni Chasaide. The results of Gobl & Ni Chasaide have been excluded because

[4] note that this list contains research done concerning TTS systems and the recognition of synthesised emotional speech using modified prosodic attributes, references in section 5.1 with the exception of Tao et al. and Mozziconacci do not use TTS

Character: Dot							
Emotion	**Max F0**	**Mean F0**	**F0 Range**	**Max Int**	**Mean Int**	**Int Range**	**Speech Rate**
Fear	1.25	1.20	1.22	1.02	1.02	1.13	1.19
Joy	1.37	1.17	1.80	1.04	1.04	1.06	1.12
Anger	1.45	1.30	1.77	1.05	1.04	1.38	1.21
Sadness	1.20	1.25	1.03	1.04	1.02	1.23	1.20
Surprise	1.24	0.99	1.62	1.01	0.98	1.25	1.08
Character: Magda							
Emotion	**Max F0**	**Mean F0**	**F0 Range**	**Max Int**	**Mean Int**	**Int Range**	**Speech Rate**
Fear	0.88	1.06	0.61	0.95	0.99	0.76	0.78
Joy	1.12	0.98	1.37	1.00	0.98	0.94	0.75
Anger	1.33	1.68	1.07	1.05	1.05	1.16	0.91
Surprise	0.89	0.89	0.95	0.93	0.94	0.73	0.96
Character: Frederik							
Emotion	**Max F0**	**Mean F0**	**F0 Range**	**Max Int**	**Mean Int**	**Int Range**	**Speech Rate**
Anger	0.90	0.97	0.83	1.02	1.03	0.96	1.14
Sadness	1.56	1.52	1.69	1.03	1.03	1.04	1.43
Surprise	1.46	1.32	2.16	1.04	1.05	1.16	1.27
Character: Tante							
Emotion	**Max F0**	**Mean F0**	**F0 Range**	**Max Int**	**Mean Int**	**Int Range**	**Speech Rate**
Anger	1.29	1.31	1.36	1.04	1.06	1.25	1.04

Table 5.6: Linear relation between emotions and neutral for all characters

they focused only on voice quality and Festival cannot influence voice quality. The tables with the prosody rules were then summarised to produce - per emotion - a table with the range of the following measured items: Mean F0, F0 Range, Tempo, Loudness and an ”other” category. Because of the different units used in most of the publications, the measurements that do not use compatible measurement units can only be compared on a qualitative basis.

Tables 5.7 to 5.11 show the summaries of the prosody rules for synthesising emotional speech in literature. The quantitative figures represent the various figures used in the literature that had comparable units, and the overall assessment contains a qualitative conclusion by me. The scale used for the qualitative terms is as follows (from very negative to very positive): Much lower, Lower, Decreased, Normal, Increased, Higher, Much higher. This scale was used in order to be able to compare all the different units used in all the prosody rulesets.

Happy/Joy	**Quantitative figures**	**Overall assessment**
F0 Mean	-30%..+50%	Varied, bigger tendency to go up
F0 Range	slightly increased’ to ’Much higher’ (+138%)	Overall usage varies between higher and much higher
Tempo	-20%..+30%	Increased (around 20%) mainly caused by pause duration decrease
Loudness	+0..+3dB	Increased, more variance
Other		Stressed syllables F0 higher

Table 5.7: Summary of happiness prosody rules

Sadness	**Quantitative figures**	**Overall assessment**
F0 Mean	+37Hz..-30Hz	Decreased
F0 Range	+40%..-50%	Lower
Tempo	-8%..-40%	Lower
Loudness	-5%..-30%	Decreased
Other		Articulation precision down, Jitter up

Table 5.8: Summary of sadness prosody rules

Anger	**Quantitative figures**	**Overall assessment**
F0 Mean	-50%..+69%	Varies, cold anger: decreased, hot anger: higher
F0 Range	+12.5%..+100%	Higher
Tempo	+2%..+50%	Higher
Loudness	+10%..+6dB	Higher
Other		Articulation precision up (especially for stressed syllables)

Table 5.9: Summary of anger prosody rules

Fear	**Quantitative figures**	**Overall assessment**
F0 Mean	+5%..+200%	Much higher
F0 Range	-5%..+100%	Increased
Tempo	+12%..+50%	Higher (probably depending on emotion strength)
Loudness	+10..+20%	Increased, low variance
Other		Stressed syllables F0 up

Table 5.10: Summary of fear prosody rules

Surprise	**Quantitative figures**	**Overall assessment**
F0 Mean	+0%..+15%	Higher
F0 Range	+15%..+35%	Increased
Tempo		Increased
Loudness	+3dB..+5dB	Increased

Table 5.11: Summary of surprise prosody rules

The raw acoustic measurements from the fragments have been converted to a linear relation. This linear relation is, per character, a comparison between the values of each fragment (except the 'None' emotional fragments), and the average of the values for the 'None' emotional fragments of that same character. The resulting linear factors have been translated to a qualitative statement using the following 7-point scale (from low to high):

- Much lower (− − −): factor is smaller than 0.5 (-50% and down)
- Lower (−−): factor is between 0.8 and 0.5 (-20% to -50%)
- Decreased (−): factor is between 0.95 and 0.8 (-5% to -20%)
- Normal (o): factor is between 0.95 and 1.05 (5% variation with regard to the neutral average[5]
- Increased (+): factor is between 1.05 and 1.25 (+5% to +25%)
- Higher (++): factor is between 1.25 and 1.6 (+25% to +60%)
- Much higher (+++): factor is bigger than 1.6 (+60% and up)

This scale is not symmetric with regard to neutrality. This is because human perception is logarithmic. For example, we notice small frequency changes a lot better when they are changes around a low frequency (40Hz with changes of 5Hz). However, at higher frequencies, small changes

[5] The 5% is a margin for measurement error. Also the minimum values in the summary changes above differ from the relative norm by an amount of 5% or more, with the exception of 3 specific measurements)

go unnoticed (15000Hz with changes of 5Hz). The asymmetry of the above scales is an attempt to roughly model this logarithmic perception into the qualitative perceptive categories ([43] chapter 9.5).

The qualitative measurements gained this way were compared with the qualitative results from the literature based on matching emotions. The result can be found in tables 5.12 to 5.15.

Emotion	Fear	Fear	Fear	Happiness	Anger
Mean F0	+	+	++	++	++
F0 Range	−−	−	−−	+	+
Mean Int	o	+	o	+	+
Speech Rate	o	+	+	o	+
Check	F0	F0,Int, SR	F0,SR	F0,F0R, Int	F0,F0R, Int,SR

Table 5.12: Qualitative comparison of fragments by character Dot, part 1

Emotion	Sadness	Sadness	Sadness	Surprise	Surprise
Mean F0	++	++	+	−	+
F0 Range	−−	−−	−−	++	−−
Mean Int	o	o	+	o	o
Speech Rate	++	o	−	+	-
Check	F0R	F0R	F0R,SR	F0R,SR	F0

Table 5.13: Qualitative comparison of fragments by character Dot, part 2

Emotion	Fear	Happiness	Anger	Surprise
Max F0	−	+	++	−
Mean F0	+	o	+++	−
F0 Range	−−	++	+	o
Mean Int	o	o	+	−
Speech Rate	−−	−−	−	o
Check	F0	F0,F0R	F0,F0R, Int	

Table 5.14: Qualitative comparison of fragments by character: Magda

Emotion	Anger	Sadness	Surprise	Surprise	Anger
Char	Frederik	Frederik	Frederik	Frederik	Tante
Max F0	−	++	+	+++	++
Mean F0	o	++	+	++	++
F0 Range	−	+++	++	+++	++
Mean Int	o	o	+	o	+
Speech Rate	+	++	+++	−−	o
Check	SR		F0,F0R, Int,SR	F0,F0R	F0,F0R, Int

Table 5.15: Qualitative comparison of fragments by characters: Frederik and Tante

The "Check" row lists the attributes which match on direction (increase, decrease, no change) with 'F0' representing Mean F0, 'F0R' representing F0 Range, 'Int' representing the mean intensity (loudness) and 'SR' representing the speech rate.

5.4 Discussion

5.4.1 Quality and quantity of fragments

While tables 5.12 to 5.15 show that almost all (17 of the 19) fragments matched the prosody rules from literature in a qualitative view on one or more measured attributes - which is not really all that difficult considering the fact that 4 of the 5 emotions (happiness, anger, fear and surprise) are linked to increases in F0, F0 range, loudness and rate of speech - the absolute measurements are quite diverse, but so are the different prosody rule-sets found in literature.

The surprisingly low number of matches for loudness may be the result of volume normalisation done in the recording studios, which would have decreased the amount of variance in the entire story, and thus making discerning emotions by loudness only possible when extremes are involved. The low amount of matches for speech rate could have its origins in the fact that these are child-oriented fairy-tales. As such, storytellers can not start talking too fast, risking the audience (children) risk being unable to keep up. It can be attributed to a general story-telling style, but in order to verify that this style overrides speech rate changes associated with certain emotions, more research on this specific subject is required.

The number of fragments (12 for the character 'Dot', 7 for character 'Magda', 7 for character 'Frederik', and 4 for character 'Tante') is too low to be of much statistical use. Especially considering that those fragments are divided over various emotions (6, 5, 5 and 2 for the characters 'Dot', 'Magda', 'Frederik' and 'Tante' respectively).

Considering the low Krippendorff's alpha value (0.284), which indicates that the people who had to link the fragments to an emotion were quite poorly in agreement, it can be argued that the emotions were not always present or as clearly present as hoped for, which caused the fragments to be improperly labelled. In order to avoid emotion labelling issues like this in the future, another approach on how to categorise emotions is discussed in the section 6.1.

The various measurements performed on the fragments did not show recognisable patterns on any of the attributes that were targeted for use in the TTS system (i.e. those that the TTS system can influence, being: Mean F0, Max F0, F0 Range, Mean Intensity, Max Intensity, and Speech Rate). This could be caused by the emotion labelling, which would cause measurements to be counted towards the 'wrong' emotion. It could also be caused by the fact that the same emotion can be expressed in different ways (with different prosodic configurations).

The per-word analysis, in combination with a grammar parser, could shed some more light on whether an emotional charge is focused on specific words, and to what degree. For example, the prosody rules by Burkhardt & Sendlmeier [4] and Cahn [5] mention the influence of the F0 of stressed syllables, the rate of speech for specific words, and the amount of - and speed of - pauses, but it is my opinion that it is too time-consuming to research these effects in detail at this moment, especially with the current set of fragments.

The origin of the diversity of the measured acoustic correlates of the emotional speech fragments could lie with the source selected for the speech fragments. If an actor was employed to create emotional speech fragments containing a targeted emotion, the whole process of emotion labelling - which lead to quite disagreeing (according to Krippendorff's alpha value) results - would not have been necessary. It would also create the opportunity to have multiple speech fragments containing the same text but different emotions. This would make the acoustic analysis easier because the utterances contain the same words in the same order. The acoustic analysis was performed measuring a lot of attributes, most of which were of no use at all. It would have been better to focus on the parameters which are currently modifiable by the TTS system. There are also parameters which the TTS system currently does not allow to be modified, but which can be modified if the appropriate changes are made to the TTS system. These parameters, such as pause- and vowel duration, should have been included in the acoustic analysis.

All in all, the measurements on their own are too few and too incoherent to create a model that could be expected to yield good results, so in order to create a model that is not a complete stab in the dark, I decided to abandon the measurements in the previous section and use the common elements of the most well recognised rule-sets found in literature as a basis for the model in this research. These models were not all based on (and tested with) the Dutch language, which is why this was a backup for the presence experiment, but at the moment it is the best data available.

This comes down to the rules shown in table 5.16. This is a table of the prosody modification rules which are recognised best in a perception test, as displayed in Schröder [46] pg. 94.

Emotion Study Language Recognition	Parameter settings
Joy Burkhardt & Sendlmeier [4] German 81% (1/9)	Mean F0: +50% F0 range: +100% Tempo: +30%
Sadness Cahn [5] American English 91% (1/6)	Mean F0: ”0”, reference line ”-1”, less final lowering ”-5” F0 range: ”-5”, steeper accent shape ”+6” Tempo: ”-10”, more fluent pauses ”+5”, hesitation pauses ”+10” Loudness ”-5”
Anger Murray & Arnott [34] British English	Mean F0: +10Hz F0 range: +9 s.t. Tempo: +30 wpm Loudness: +6dB
Fear Burkhardt & Sendlmeier[4] German 52% (1/9)	Mean F0: +150% F0 range: +20% Tempo: +20%
Surprise Cahn [5] American English 44% (1/6)	Mean F0: ”0”, reference line ”-8” F0 range: ”+8”, steeply rising contour slope ”+10”, steeper accent shape ”+5” Tempo: ”+4”, less fluent pauses ”-5”, hesitation pauses ”-10” Loudness: ”+5”
Boredom Mozziconacci [30] Dutch 94% (1/7)	Mean F0: End frequency 65Hz (male speech) F0 range: excursion size 4 s.t. Tempo: duration relative to neutrality: 150%

Table 5.16: The model: best recognised prosody rules from literature [46] pg. 94

5.5 Conclusion

As discussed above, the number of fragments used for the acoustical analysis was way too small to provide a solid statistical base to build a model upon. These few fragments, with possible incorrect emotion labelling and, in retrospect, a labelling method which turned out to be too coarse, resulted in, per emotion, too erratic values (also when compared with measurements found in literature) which made it impossible to devise a model that would translate an emotional term into a well-defined set of prosodic attribute changes. This lead me to the drastic action of dropping the fragments and the results of the analysis thereof, and adopt a new approach. This approach is explained in the next chapter.

Chapter 6

Plan B

Though the results of the fragment analysis could not be used in the construction of a model that morphs neutral speech into emotional speech, there is an alternative approach. This approach was found when looking for literature that would explain the conflicting measurements of the previous chapter.

This chapter covers the introduction of another way of describing emotions: the Emotion Dimension Framework and the method in which the EDF is linked to the emotional terms used earlier in this thesis is described in section 6.1. This method and the EDF, together with an acoustical analysis performed on the Belfast database of spontaneous emotional speech by Schröder resulted in a new model (hereafter referred to as model B). This model is explained in section 6.2.

6.1 The Emotion Dimension Framework

There are other methods of describing emotions than just by using emotional labels. Schröder [46] devotes a chapter in his thesis to a literature research into the various frameworks for emotions that people have used: Emotion categories, Prototype descriptions, Physiology-based descriptions, Appraisal-based descriptions, Circumplex models, and Emotion dimensions. From these frameworks, Schröder chose to use the Emotion Dimension Framework to describe emotions in his thesis with the following remark [46] pg. 32:

> I tend to agree with Lazarus in that emotion dimensions are a reduced account of emotions, in which many important aspects of emotions are ignored, such as eliciting conditions and specific action tendencies. Nevertheless, I consider a dimensional description as a useful representation, capturing those aspects that appear as conceptually most important in many different methodologies, and providing a means for measuring similarity between emotional states. In particular, a dimensional description is particularly well-suited for . . . generating a voice prosody which is *compatible* with an emotional state expressed through a different channel, rather than fully *defining* that emotional state. A reduced description is sufficient for this purpose, and does not preclude being used in conjunction with richer descriptions.

From the remark, we can conclude that the EDF does not completely describe emotions, but does sufficiently describe emotions in order to generate voice prosody which is compatible with an emotional state. This also applies to this thesis, where we also want to generate voice prosody to allow an utterance sound emotional.

An EDF uses multiple dimensions to represent emotional states, depending on the researcher, these axes of the dimensions and their scales may vary. Schröder has investigated these different emotion frameworks in [46] Chapter 2, with the EDFs in section 2.6 specifically. These details will not be reproduced here; the interested reader is invited to read Schröder's thesis. Of all the frameworks available, he chose to use a three-dimensional EDF with the following axes:

- activation (also known as activity or arousal) ranging from passive to active.
- evaluation (also called pleasure or valence) ranging from negative to positive.
- power (also known as dominance or control)

Of these axes, the activation and evaluation dimensions are used in the same way as in the Feeltrace [10] tool. This tool can be used to track the activation and evaluation dimensions of a perceived emotion over time while listening to audiodata (which can be speech, music or other audio that can be perceived as having emotional content). About the power axis, Schröder says this ([46] p.105):

> The rich characterisation of these emotion words obtained in previous experiments [9] allows the addition of the power dimension, as associated with the emotion word, to each clip. Thus each clip is positioned on the three dimensions, with activation and evaluation changing over time during the clip and power remaining static.

However, even after carefully reading [9], it remains unclear how *exactly* these power values have been associated with the emotion words, shown in table 6.1.

In order to link the activation and evaluation axes to emotion word, Schröder used Feeltrace in [46] pg. 131. By having people who were trained in the use of Feeltrace assign both an emotion word and Feeltrace activation- and evaluation dimensions to audio fragments, the emotion word became linked to activation and evaluation values.

This resulted in table 6.1. The values are on a scale from -100 to 100.

Emotion	**Activation**	**Evaluation**	**Power**
neutral	1.8	-1.7	0
bored	-6.8	-17.9	-55.3
disappointed	2.4	-24.9	-37.2
sad	-17.2	-40.1	-52.4
worried	4.6	-26.3	-62.3
afraid	14.8	-44.4	-79.4
angry	34.0	-35.6	-33.7
interested	16.8	16.6	-6.1
excited	36.1	30.5	-5.8
loving	1.2	33.3	14.9
affectionate	0.7	37.3	21.4
pleased	19.0	38.6	51.9
confident	13.8	14.1	32.9
happy	17.3	42.2	12.5
amused	23.4	16.8	-5.0
content	-14.9	33.1	12.2
relaxed	-18.5	25.7	-5.2

Table 6.1: Positions on the three emotion dimensions for some emotion words

6.2 A new model

Schröder performed an acoustic analysis on the Belfast database of spontaneous emotional speech in [46] chapter 11. He used this analysis to extract correlations between a number of acoustic variables like F0 median, F0 range, pause duration, intensity median, intensity range, etc and the three emotion dimensions: activation, evaluation and power. These linear regression coefficients have been copied from [46] table 11.7 into table 6.2. The table only shows the values for female speech. Schröder also extracted data for male speech, but it is not necessary for the understanding of the model to have both tables available here.

As an example, Schröder calculates the angry-sounding F0 median as the result of the sum of the base value and, for each emotion dimension, the product of the regression coefficient for the

	Acoustic variable	**Unit**	**Linear Regression Coefficients**			
			Const.	**Activation**	**Evaluation**	**Power**
fundamental frequency	F0 median	Hz	200.1	**0.370**	**-0.0523**	**-0.190**
	F0 range	Hz	28.45	**0.243**	**-0.0531**	
	med. magn. F0 rises	Hz	14.57	**0.107**	(-0.0140)	
	range magn. F0 rises	Hz	20.94	**0.212**	**-0.0352**	
	med. magn. F0 falls	Hz	19.46	**0.177**	**-0.0488**	
	range magn. F0 falls	Hz	24.68	**0.244**	**-0.0727**	
	med. dur. F0 rises	sec	0.2140	**0.000379**	**0.000258**	0.000216
	rng. dur. F0 rises	sec	0.1914	**0.000305**	(-0.000164)	(0.000180)
	med. dur. F0 falls	sec	0.2448	**0.000537**	(0.000140)	(0.000173)
	rng. dur. F0 falls	sec	0.1924	**0.000311**	**-0.000309**	**0.000338**
	med. slope F0 rises	Hz/sec	76.35	**0.403**	**-0.166**	**-0.106**
	med. slope F0 falls	Hz/sec	85.85	**0.516**	**-0.191**	**-0.101**
	F0 rises p. sec.	1/sec	2.712	**-0.00593**	-0.00217	(-0.00202)
	F0 falls p. sec.	1/sec	2.486	**-0.00578**	**-0.00273**	
tempo	duration pauses	sec	0.4367	**-0.00122**	(0.0003322)	**-0.000775**
	'tune' duration	sec	1.424	**0.00355**	(-0.00105)	
	intensity peaks p. sec.	1/sec	4.090			
	fricat. bursts p. sec.	1/sec	1.410	**-0.00464**		0.00247
intens.	intensity median	cB	531.2	0.0513	**-0.0667**	0.0615
	intensity range	cB	103.2	**0.149**		
	dynamics at peaks	cB	25.89	**0.0525**	**-0.0256**	
voice quality	spectral slope non-fric.	db/oct	-7.396	**0.0147**	**-0.00293**	0.00181
	Hamm. 'effort'	-	32.98	**0.0229**		**0.00898**
	Hamm. 'breathy'	-	8.084	**-0.0235**		
	Hamm. 'head'	-	24.68	**-0.0121**		**-0.0282**
	Hamm. 'coarse'	-	16.38	**-0.0178**		**-0.0114**
	Hamm. 'unstable'	-	8.297	**0.00569**		**-0.0170**

Table 6.2: Significant linear regression coefficients for acoustic variables predicted by emotion dimensions for **female** speech. Coefficients printed in bold are significant at $p < .001$; coefficients printed as plain text are significant at $p < .01$; and coefficients printed in brackets are significant at $p < .05$.

emotion dimension with the value of the emotion dimension associated with the emotion category 'angry'. This is done with the formula 6.1. The formula is used for F0, but can be used for any of the acoustic correlates from table 6.2.

$$F0(emotion) = F0_{base} + \sum_{\forall I \in \{A,E,P\}} LRC(I) * ValueOf(I, emotion) \tag{6.1}$$

Here LRC stands for Linear Regression Coefficient, and the ValueOf function performs the table lookup that translates the emotion word into the 3-dimension values (A)ctivation, (E)valuation and (P)ower. The LRC values are taken from table 6.2.

With this formula and table 6.2, it is possible to calculate the emotional variant of each of the mentioned acoustic variables in the table for each of the emotion terms used in table 6.1.

6.2.1 The link with the Virtual Storyteller

Not all the researched emotions are present in the VST system. Currently, the it uses a list of emotions derived from the OCC model [36]. The emotions used are listed in research done by Rensen on the Virtual Storyteller in [42]. The list of emotions is divided between positive and negative emotions, each pair consists of two opposing emotions:

- “Happy for” vs “Pity”
- “Gloating” vs “Resentment”
- “Hope (for self & others)” vs “Fear (for self & others)”
- “Joy” vs “Distress”
- “Pride” vs “Shame”
- “Admiration” vs “Reproach”
- “Love” vs “Hate”

These emotions were chosen with regard to agent decisions on how to act. The use of emotions in the narration of the story is new, and not all the emotions used in the VST system are used as often as, for example the big six (Anger, Disgust, Fear, Joy, Sadness, Surprise). The reverse is also true, not all of the often used emotions are present in the VST system, for example sadness and surprise are missing. The fact that most of the emotion terms used in the VST system were not researched with respect to TTS prosody rules, means that there does not exist a model yet to morph a neutral voice into an emotional voice for those emotions.

In order to create a model that produces the best expected results, the choice was made to use the prosody rules in table tables:themodel for those emotions, and to use the transformation using Schröder’s formula and acoustic correlation data in table 6.2 for all the other emotion terms - as far as they are present in table 6.1. Emotion terms that do not have an existing model, or are not in table 6.1 can not be synthesised until the relation between the emotion word and the values on the activation, evaluation and power dimensions can be determined.

6.3 The new model

The model now uses 2 methods of changing certain prosodic attributes of a TTS voice based on the input emotions:

- If an emotion category has a prosody rule-set, the rule-set is used.
- Otherwise, the emotion category is translated into the dimensional model and the three dimensions are then used to determine the prosodic attributes that need to be altered, and the amount they need to be altered with.

This rule-set is presented in a graphical form in figure 6.1.

Table 6.3 contains the changes the model will apply to the F0, F0 Range, Speech Rate, and Intensity. The model contains other emotions (namely those of table 6.1) but the emotions in table 6.3 were the ones used in the evaluation experiment and are therefore explicitly shown here. The prosodic changes for the emotions: joy, sadness, anger, fear, surprise and boredom were derived from table 5.16. The prosodic changes for the emotion love were calculated using Schröder’s data.

Emotion	**F0**	**F0 Range**	**Speech Rate**	**Intensity**
Joy	+50%	+100%	+30%	
Sadness		-25%	-25%	-50%
Anger	+10Hz	+9st		+100%
Fear	+150%	+20%	+20%	
Surprise		+80%	+40%	+25%
Boredom		-4st	67%	
Love	-1.78Hz	-3.30Hz		-2.15%

Table 6.3: The model’s changes on prosodic parameters

6.4 Conclusion

The new approach uses another emotion description framework, namely by using three scales per emotion: activation, evaluation and power. Each of these scales has a range of -100 to 100 and is

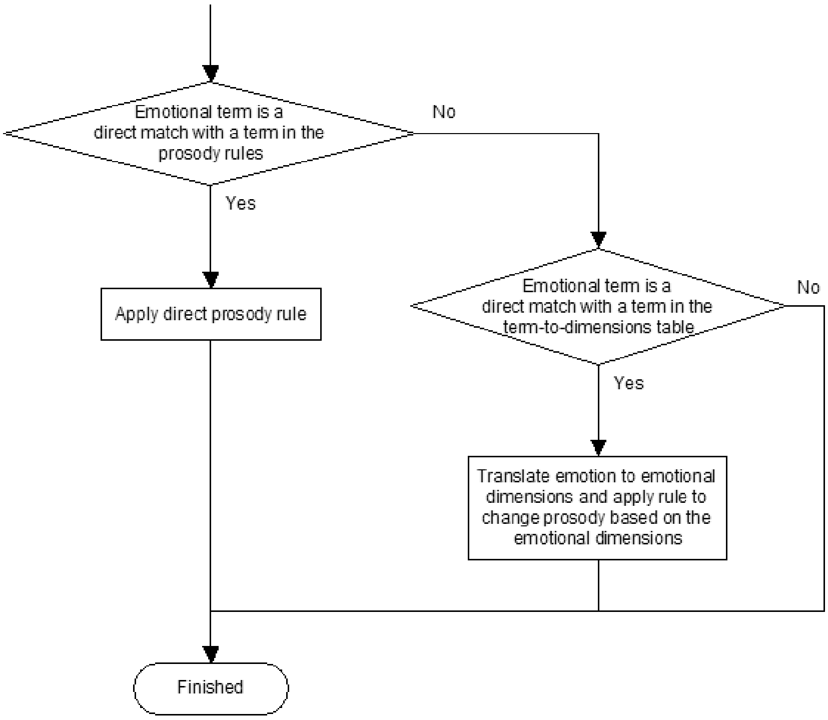

Emotional term is a direct match with a term in the prosody rules
No
Yes
Apply direct prosody rule
Emotional term is a direct match with a term in the term-to-dimensions table
No
Yes
Translate emotion to emotional dimensions and apply rule to change prosody based on the emotional dimensions
Finished

Figure 6.1: Flowchart of the model

used in conjunction with the acoustic analysis of emotional speech performed by Schröder where the values on these scales are linked to various prosodic (acoustic) attributes. The application Feeltrace [10] can be used to determine the emotional activation and evaluation values of audio fragments. When someone also annotates the same audio fragments with emotional terms, the emotional term becomes linked to the activation and evaluation values assigned to the fragment using Feeltrace. This provides a good method for adding other emotional terms to the list of supported emotional terms by model B. However, the origins of the values on the power scale provided by Schröder in table 6.1 is still unclear, preventing the addition of power values for other emotional terms.

Model B uses existing - and tested - prosody rules to morph the neutral (default) TTS voice into an emotional one for the emotions joy, sadness, anger, fear, surprise and boredom. These prosody rules have been tested by the researchers who created these rules, and scored an acceptable recognition rate on perception tests as shown in table 5.16. The emotions that do not have existing prosody rules are morphed using the three-dimensional EDF and Schröder's acoustic correlates in tables 6.1 and 6.2. The process that model B follows is shown in figure 6.1.

Chapter 7

Implementation

In this chapter, the process of implementing the model of the previous chapter in the TTS engine Festival will be detailed. All prerequisites that are unmet in the Festival engine will also be discussed, as well as the requirements to overcome these dependencies. Also, the process of implementing the prior work by Koen Meijs [27] in Festival-Nextens will be covered in this chapter.

7.1 Narrative model

Koen Meijs has performed an analysis of the narrative speaker style in [27] by comparing the use of prosody between a storyteller and someone reading the news. This analysis resulted in the following items:

- A narrative speaking style. This focuses on the enhancements of accented syllables of key words in sentences. These accents are annotated in the input with <sentence_accent>-tags. Accents are enhanced by adding a 'bump' to the speakers normal F0. This bump is defined by the formula:

$$Target\_Pitch = Current\_Pitch * (1 + sin(\pi * t)/(Average\_Pitch/Target\_Pitch\_Increase)) \tag{7.1}$$

  with t=0,25..0,75 and the pitch variables are in Hz. The Target_Pitch_Increase variable has been set by Meijs to 30Hz. The intensity of accents is increased by 2dB and the duration of the vowels of the accented syllable is multiplied by a factor 1,5. The global speech rate for the narrative speaking style was determined to be 3,6 syllables per second.
- The sudden climax is the first of two climax types specified in Meijs' research. The sudden climax changes the pitch, intensity and the duration of a part of a certain utterance. The start and end of the sudden climax have to be annotated in the input. The pitch is increased in a linear fashion, increasing the pitch by 80Hz at the end of the climax. After the sudden climax, the next syllable starts at the same frequency as the climax ended with. The intensity of the speech gets a 6dB boost starting at the beginning of the sudden climax and is decreased back to normal during the course of the climax. The speech rate remains unchanged.
- The second climax type is the increasing climax. This climax is an extended version of the sudden climax. It has a starting- and ending position just like the sudden climax, but also a climax_top indicator. The pitch of the increasing climax gets a boost of 25Hz at the start, and then a pitch 'bump', just like the narrative speaking style accent, of maximally 35Hz is added to the pitch in between the start and the end of the climax, with the maximum lying at the designated top of the climax. The intensity remains unchanged during an increasing climax. The duration of accented vowels of words that have sentence accent are multiplied by a factor that is gradually increased from 1 at the start of the climax to 1.5 at the top. At the top of the climax, a pause of 1,04s is inserted. After the top, the duration of accented vowels is gradually reduced back to normal.

The values in this list are taken from [27] section 7.4.

7.2 Input format

The current narrator implementation requires that its input is formatted using SSML. In the implementation created by Meijs, this markup language has been extended with certain tags to allow the indication of the aforementioned narrative speaking style and climax types. It is my opinion that changing SSML this way is not the proper method. An intermediary XML format can be used to indicate these styles just as well, and this intermediary XML can then be translated into SSML. During this translation process, the changes in pitch, speech rate and intensity dictated by the narration rules of Meijs can be translated into properly placed <prosody> tags, which are already in the SSML specification. This leads to the data flowchart depicted in figure 7.1.

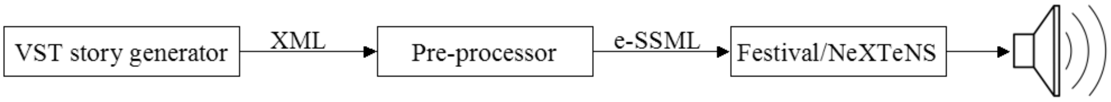


Figure 7.1: System data flowchart

The SSML specification is still limited in respect to other prosodic attributes such as proximity (to the speaker), shouting indicators, breathiness, creakiness, articulation, ToBi, attitudes and emotions. There is an extension to SSML, which I will refer to as e-SSML in this thesis, which attempts to resolve this limitation in [14]. However, the emotion attribute suggested in [14] is limited to emotion terms which is not compatible with the emotion dimension framework adopted in model B. To this end, the emotion attribute is included in the intermediary XML format, and is then automatically translated into the appropriate changes in pitch, speech rate and intensity that SSML already supports.

The intermediary XML format supports the following tags (closing tags are not listed):

- <sentence_accent> with attribute ”extend”, which either has the value ”yes” or ”no”. The ”extend” attribute signals if the duration multiplier needs to be applied or not. The sentence_accent tag is placed around a single word.[1]
- <climax> with the attribute ’type’, which either has the value ”immediate” or ”increasing”. The ”immediate” value signifies a sudden climax, the ”increasing” value signifies an increasing climax. The climax tag is placed around one or more words (in case of the immediate type) or parts of one sentence (in case of the increasing type).
- In case of an increasing climax type, the empty <climax_top/> tag signals that the top of the climax has been reached, and the F0 contour (which has been rising until now) can decrease again.[2]
- <quote> with attributes ’character’ and ’emotion’. The value of the character attribute is the name of the character uttering the quote. The value of the emotion tag is a literal string of the emotion (in English).

The language generation module in the VST will need to add Meijs’ narrative tags, as well as the <quote> tags to the plain text of the story. At the moment this is not implemented yet; the language generation module produces plain text only. The XML data will then be translated into e-SSML where the sentence accents, climaxes and quotes will be replaced by prosody tags with the appropriate attributes. This translation process from XML to e-SSML is done by an application hereafter referred to as the ”pre-processor”:

[1] This is different from the implementation by Meijs, which specified the tag to be placed around a single syllable. Because placing a tag inside a word breaks the Festival intonation algorithm, this has been changed.

[2] This is different than used by Meijs, who put the tag in front of the word that was at the top of the climax, but I feel that putting it right after the top is a more logical position because it signals the start of the pitch decline.

- A sentence accent tag will be replaced by a prosody tag with contour and rate attributes.
- An increasing climax tag will be replaced by prosody tags with rate, volume and pitch attributes. This is because the speech rate changes over time, so separate tags are required.
- An immediate climax tag will be replaced by prosody tags with pitch and volume attributes.
- For quote tags, there are two possible methods. This is because the e-SSML specification allows an "emotion" attribute in a prosody tag. This provides the option of simply passing the desired emotion term on to the TTS system, instead of converting the emotion term into a set of appropriate prosody tags. The tables used in transforming an emotionally neutral utterance into an utterance with a different emotion, by containing values to be used in the prosody attributes per emotion, are either present in the TTS environment or in the pre-processor. If the tables are present in the TTS environment, the quote tags can easily be replaced by prosody tags with only an emotion attribute, leaving the actual work to the TTS system. If the tables are present in the pre-processor, the quote tags will be replaced by prosody tags with pitch, rate, breathiness, creakiness, articulation, volume and contour attributes, configured using the emotion translation tables. It is easier to put the emotion translation tables in the pre-processor, because this requires less tampering with the TTS environment. The only requirement for the TTS system is that it can parse e-SSML and act upon it. The tables can be modified without requiring the user to have advanced knowledge on the inner working of the TTS system. It also allows for the option to use other methods of describing the target emotion than by using emotion terms. If future versions only use the emotion dimensional framework, then only changes to the pre-processor need to be made, instead of having to change the (much larger) TTS system again.

As an example, a VST XML file is shown below, as well as the translated SSML file. They contain a short example story in Dutch:

```
Er was eens een verhaal. Dit was niet <sentence_accent extend="no">zomaar
</sentence_accent> een verhaal, nee, het was een <sentence_accent extend="yes">
heel</sentence_accent> spannend verhaal.
Jantje liep in het bos. Het was een groot donker bos en in de verte kon hij een
wolf horen huilen.
De wind waaide door de boomtoppen die hierdoor woest heen en weer zwiepten.
<quote character="name" emotion="fear">Misschien was dit toch niet zo'n goed
idee.</quote> zei hij tegen zichzelf.
<climax type="increasing">Jantje begon naar de rand van het bos te lopen maar
<climax_top/> toen sloeg de bliksem plotseling vlakbij in.</climax>.
```

Example story in VST XML

```
<?xml version="1.0" encoding="ISO-8859-1"?>
<!DOCTYPE speak PUBLIC "-//W3C//DTD SYNTHESIS 1.0//EN" "synthesis.dtd">
<speak version="1.0" xmlns="http://www.w3.org/2001/10/synthesis"
xmlns:xsi="http://www.w3.org/2001/XMLSchema-instance" xml:lang="nl">
Er was eens een verhaal. Dit was niet <prosody contour="(0%,+28.280Hz)
(25%,+36.960Hz) (50%,+40Hz) (75%,+36.960Hz) (100%,+28.280Hz)">zomaar</prosody>
een verhaal, nee, het was een <sentence_accent extend="yes">heel
</sentence_accent> spannend verhaal.
Jantje liep in het bos. Het was een groot donker bos en in de verte kon hij
een wolf horen huilen.
De wind waaide door de boomtoppen die hierdoor woest heen en weer zwiepten.
<prosody pitch="150%" range="+20%" rate="+20%">Misschien was dit toch niet zo'n
goed idee. zei hij tegen zichzelf.</prosody>
<prosody volume="+10"><prosody pitch="+25.000Hz" rate=".952">Jantje</prosody>
<prosody pitch="+28.500Hz" rate=".909">begon</prosody>
<prosody pitch="+32.000Hz" rate=".869">naar</prosody>
<prosody pitch="+35.500Hz" rate=".833">de</prosody>
```

```
<prosody pitch="+39.000Hz" rate=".800">rand</prosody>
<prosody pitch="+42.500Hz" rate=".769">van</prosody>
<prosody pitch="+46.000Hz" rate=".740">het</prosody>
<prosody pitch="+49.500Hz" rate=".714">bos</prosody>
<prosody pitch="+53.000Hz" rate=".689">te</prosody>
<prosody pitch="+56.500Hz" rate=".666">lopen</prosody>
<prosody pitch="+60.000Hz" rate=".645">maar</prosody>
</prosody><break time="1040ms"/><prosody volume="+6">
<prosody pitch="+60.000Hz" rate=".666">toen</prosody>
<prosody pitch="+50.000Hz" rate=".706">sloeg</prosody>
<prosody pitch="+40.000Hz" rate=".750">de</prosody>
<prosody pitch="+30.000Hz" rate=".800">bliksem</prosody>
<prosody pitch="+20.000Hz" rate=".857">plotseling</prosody>
<prosody pitch="+10.000Hz" rate=".923">vlakbij</prosody>
<prosody pitch="+0Hz" rate="1.000">in.</prosody></prosody>.
</speak>
```

Example story in SSML

7.3 Additional Constraints

Because the TTS front-end system of Festival is designed to be used for the English language, NeXTeNS uses its own modules for accent placement, intonation- and duration modification etc. As such, the code for applying the narrative speaking style and emotions has to be interfaced with NeXTeNS' modules instead of those of Festival. The e-SSML specification does not allow speech rate changes to be defined by vowel duration and pause duration. This is used by model B for the prosody changes using Schröder's work. Schröder has correlated duration pauses and 'tune' (vowel) duration to the three dimensions of the EDF. Updating the e-SSML specification with attributes to support pause- and vowel duration, as well as implementing it in NeXTeNS was estimated to take too much time away from the main implementation. As such the pause- and vowel duration specification and implementation have been left as future work.

7.4 SSML parser implementation details

Festival uses tree structures to store data while processing the text. First only the input text is stored, but when the various modules like the tokenize-, accent placement-, intonation- and duration modules have done their jobs, their input comes from the same tree structures, and their output is saved in the tree structures as well. When parsing the SSML document, the tags are also stored in these tree structures (in their own tree) complete with all attributes and values. These SSML tags are then linked to the words, and in doing so with all the data that is related to the words, in the tree structures. This way, whenever a module is done with its work, for example the intonation module, the SSML tags can be applied by scanning the tree of SSML tags for tags with intonation modifications, and applying those tags to the words the tags are linked to. In order for the SSML tags to be applied, all modules that perform a task where an SSML tag can influence the outcome call a specific function that applies the SSML tags, but only after such a module has performed its own task first. In the end, the input text is also present in the tree structures as a list of segments, each with a start time and a duration as well as zero or more pitch targets per segment. These segments are then supplied to the MBROLA program which performs the task of selecting the proper diphones, modifying the pitch to match the pitch information that came with the segments, and eventually produce audio data with the synthetic speech.

The Maintainer manual, which can be found in appendix B, goes into greater detail about the implementation. It contains instructions on how to install Festival and NeXTeNS and a list of functions that are called when Festival has to parse SSML data. It also lists the procedures defined together with a short description of what they do.

7.5 Festival web interface

In order to facilitate the use of the implemented Festival SSML support, a number of web pages were made using PHP[3] and MySQL[4]. PHP is a scripting language especially suited for web development, and MySQL is an SQL database server. The web pages allow the user to enter a single utterance, and four prosodic attributes: F0, F0 range, Speech rate and Volume. There is also a combo-box with the seven used emotional terms (Happiness, Sadness, Anger, Fear, Surprise, Boredom and Love) which will automatically fill the prosodic attributes fields with the proper values. With a push of the button, the entered utterance, together with the prosodic attributes, is encoded into SSML and sent to the Festival TTS server. After the synthesis process is complete, the returned speech is saved on the computer where the web-server is running. This is done because the TTS process and the transfer of the speech data can take anywhere from around five seconds or more. If the user wants to listen to the same text more than once, the system will play the stored speech on the web-server, instead of having to employ the TTS process each time. Each utterance for which speech data is saved on the web-server is listed in a section on the right side of the web interface. The user can play or delete a selected file, delete all the files, and refresh the list. In order to separate the speech files from the different users, a small login function has been implemented. A user is identified by a username and the IP-address the web-server detects the connection from. This login function is not meant as a security measure, but simply to avoid user A from adding to, or deleting from, user B's speech files.

Should anything go wrong during the TTS process, for example when the text parser encounters syllables which can not be synthesised using the Dutch language (default language setting), Festival will not return an error code indicating what went wrong. This results in only very crude error detection ("Something went wrong, check everything and try again."). Future work on the Festival server mode is required in order to provide detailed error messages in case something goes wrong. A small section at the bottom of the web interface is reserved for action logging and error notification. If anything goes wrong, it will be shown here.

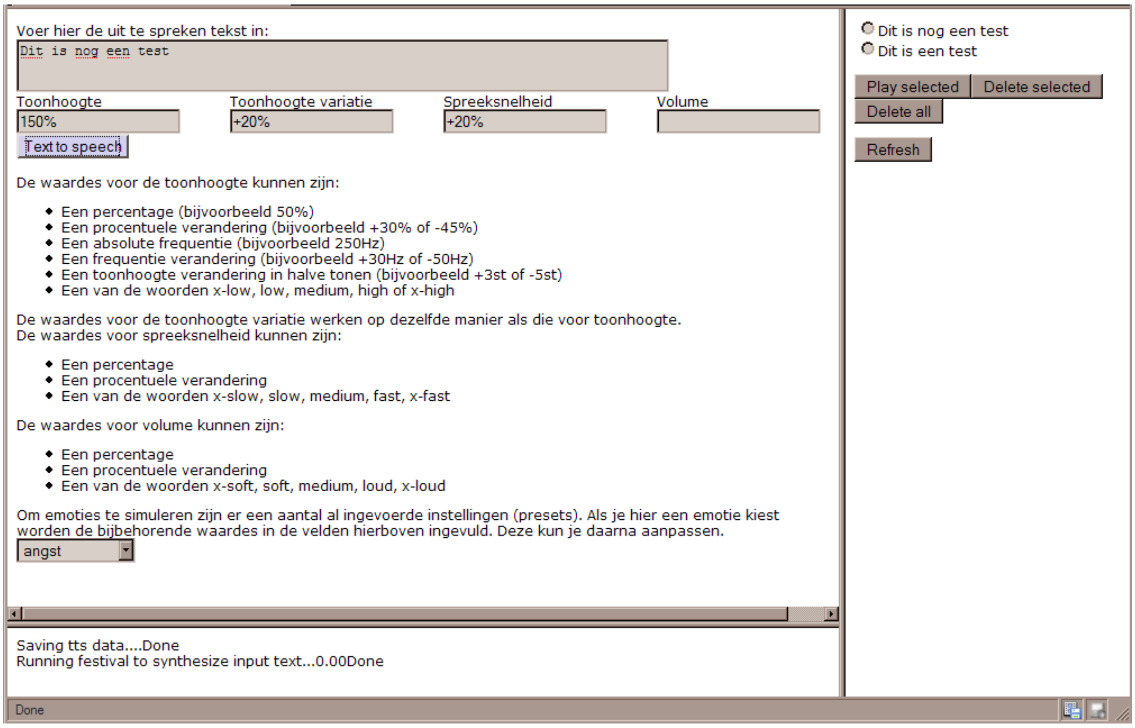
The interface is easily modifiable, making it easy to change the user input from a single utterance with prosodic attributes to, for example, a user-supplied SSML file. This file can then be sent directly to the Festival server. The resulting audio is still saved on the computer where the web-server is running.

Below is a screenshot of the main interface (figure 7.2):

The source code of the Festival web interface has been included in the digital appendix.

[3] http://www.php.net

[4] http://www.mysql.org



Voer hier de uit te spreken tekst in:

```
Dit is nog een test
```

Toonhoogte: 150%

Toonhoogte variatie: +20%

Spreeksnelheid: +20%

Volume:

De waardes voor de toonhoogte kunnen zijn:

- Een percentage (bijvoorbeeld 50%)
- Een procentuele verandering (bijvoorbeeld +30% of -45%)
- Een absolute frequentie (bijvoorbeeld 250Hz)
- Een frequentie verandering (bijvoorbeeld +30Hz of -50Hz)
- Een toonhoogte verandering in halve tonen (bijvoorbeeld +3st of -5st)
- Een van de woorden x-low, low, medium, high of x-high

De waardes voor de toonhoogte variatie werken op dezelfde manier als die voor toonhoogte.
De waardes voor spreeksnelheid kunnen zijn:

- Een percentage
- Een procentuele verandering
- Een van de woorden x-slow, slow, medium, fast, x-fast

De waardes voor volume kunnen zijn:

- Een percentage
- Een procentuele verandering
- Een van de woorden x-soft, soft, medium, loud, x-loud

Om emoties te simuleren zijn er een aantal al ingevoerde instellingen (presets). Als je hier een emotie kiest worden de bijbehorende waardes in de velden hierboven ingevuld. Deze kun je daarna aanpassen.

angst

○ Dit is nog een test
○ Dit is een test

```
Saving tts data....Done
Running festival to synthesize input text...0.00Done
```

Figure 7.2: Screenshot of the main Festival web interface

Chapter 8

Evaluation experiment

In this chapter, the design and execution of another experiment is presented. The goal of this experiment is to show if the model created in chapter 5 produces emotional utterances where the emotion is recognisable. The models to morph neutral speech to emotional speech that have been used have been tested with good results by each model's respective creator, but it is unknown if those models will still work as well in the current application. The test subjects will be presented with fragments based on sentences that have no distinct emotional semantic aspect; the meaning of the sentences does not point towards a single emotion. The subjects will then be asked to verify if the emotional content can be identified and, if not, why not. The results of this experiment can be used to update the models with different parameters in order to proceed towards a more easily recognised expression of emotion.

Given the fact that the models used were not all constructed from data retrieved from Dutch speakers, it is not clear if these models are just as effective if they were applied to the Dutch language. Also, the models that NeXTeNS uses to implement Dutch intonation contours, for example, might not produce the expected results when combined with an emotion-generating model. In order to find if and to what degree the models result in identifiable emotional utterances the evaluation test was created.

The chapter is set up as follows: First the material presented to the test subjects is discussed in section 8.1. Because the environment- and participant constraints are the same as with the emotion perception experiments described in chapter 4, sections 4.1 and 4.3 they are not duplicated here. The interface of the test is presented in section 8.2. This section details the type of question asked and the type of answers a test subject could submit. Finally the experiment results are presented and discussed in section 8.3.

8.1 Material

The goal of this test is to verify whether the model has applied the appropriate prosodic changes to a sentence in order to allow people to recognise the intended emotion. In order to do this, the speech fragments supplied to the test subject must not contain any semantical indication of the emotion used. In other words, the emotion should be recognised from the prosodic aspect of the fragment, not from the semantics or the syntax of the presented sentence. Thus the sentences will have to be either emotionally neutral or emotionally ambiguous, thereby forcing the test subject to perceive the emotion from the prosodic aspect of the fragment. To this end, the following emotionally neutral/ambiguous sentences have been used:

- "Waar heb je dat vandaan gehaald?" (Where did you get that from?)
- "Dat had je niet moeten doen." (You shouldn't have done that)
- "Morgen komt mijn schoonfamilie op bezoek." (Tomorrow my in-laws will come visit)

Each of the above sentence was morphed to each of the implemented emotions: sadness, surprise, anger, fear, love, joy, boredom. Additionally, one emotionally biased sentence was used for each of the emotions the model implements. Finding a sentence that was biased towards sadness turned out to be harder than initially expected, because almost all of those sentences could also be used to express anger. Because of this, the sentence used for anger and sadness is the same:

- Anger, sadness: ”Dat had je best zelf kunnen verzinnen.” (You could've come up with that on your own easily)
- Love: ”Wat ben je toch een schat.” (You're such a darling)
- Fear: ”Ik hoop niet dat we hem kwijt zijn geraakt.” (I hope we didn't lose him)
- Boredom: ”Wanneer komt hier nou eens een eind aan.” (When will this finally end)
- Surprise: ”Maar dat had ik niet verwacht.” (But I didn't expect that)
- Joy: ”Deze test is bijna afgelopen.” (This test is almost over)

These added sentences serve as a failsafe should the model produce quite unexpected results. If the model produces prosody that does not agree with the intended emotion (which is reinforced by the semantics of the sentence), then a negative result will be returned. On the other hand, if the prosody produced is even slightly compatible with the intended emotion, these sentences will have a very high recognition rate.

8.2 Interface

8.2.1 Introductory page

As with the presence experiment, the subject is first shown an introductory page which explains what will be presented to him, and what will be asked of him. The introductory page also provides the subject with four example fragments of synthesised speech. This is to familiarise the subject with the quality and quirks of the synthetic voice, and also to allow the subject to adjust the volume of his audio equipment to an acceptable level. Also, the amount of fragments that will be presented during the experiment are given in order to give the subject an idea on the approximate duration of the experiment.

8.2.2 Main page design

The question ”What to present to the test subject?” is closely linked to the question ”What to ask from the test subject?”. The results of the presence experiment have shown that multiple-choice emotion identification produced results with a very low agreement factor. It is because of this that I have decided that the evaluation test should use a different answering mechanism. Instead of asking the test subject to identify an emotion, the question is simplified to ”Does this sound like (emotion) X?”. Here the target emotion is explicitly provided to the test subject in order to avoid mixing two problems: incorrectly identified emotions and incorrectly rendered emotions. Because the focus here is to find out whether or not the model results in correctly rendered emotions, the elimination of the first problem is necessary. This does introduce a bias towards the target emotion which can result in more positive answers than would otherwise be the case, but it is my opinion that if the target emotion would need to be identified by the test subject, this experiment would fall to the same problems that befell the presence experiment (see section 5.5).

In an attempt to get more usable feedback than a mere ”yes” or ”no”, the test subject can indicate, if the fragment does not sound like emotion X, what is causing this. This is done by providing a six-point scale for each of the four prosodic aspects used in imitating the emotion: Pitch, Pitch Variance, Speech Rate, and Volume. The six-point scale consists of the ratings: ”way too low”, ”too low”, ”ok”, ”too high”, ”way too high”, and ”don't know” for Pitch, Speech Rate and Intensity, and the ratings: ”way too little”, ”too little”, ”ok”, ”too much”, ”way too much”, and ”don't know” for Pitch Range. The default setting for the six-point scales is ”don't know”. These feedback indicators will help in pointing out the most obvious flaws in the model.

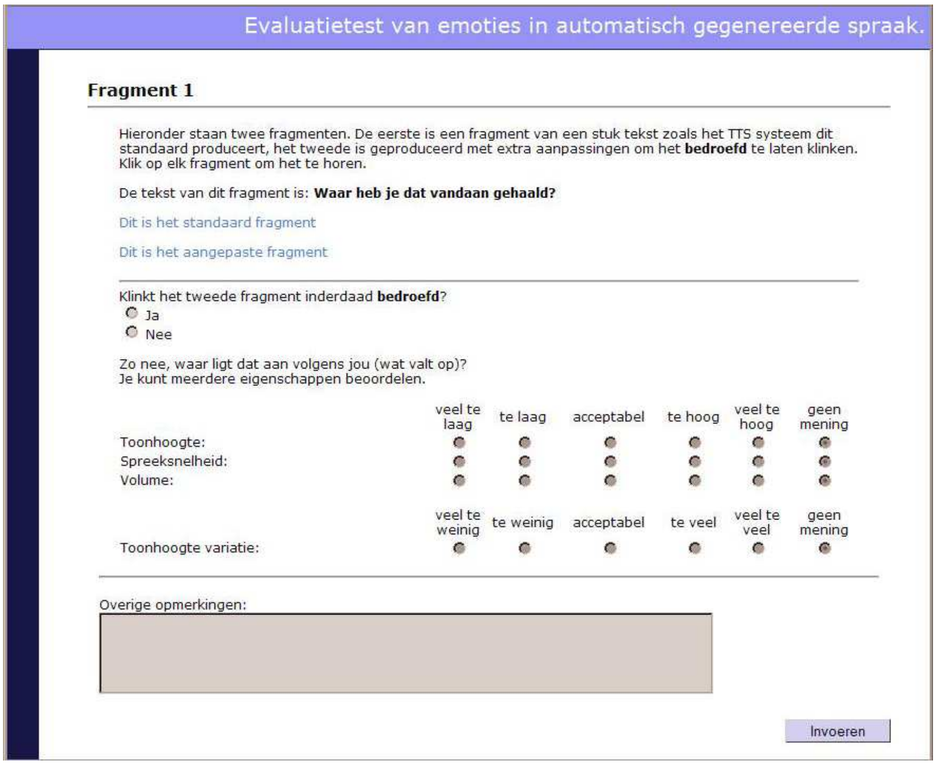
There is also a comment field, where people can leave other comments about the fragments that have not been addressed yet.

So now the question of ”What to ask” has been answered, leaving us with ”What to present”. Here a few options present themselves:

- Present a single emotional utterance.
- Present a neutral and the emotional variant of an utterance.
- Present multiple emotional utterances.

When presenting a single utterance, the test subject can only judge based on the currently presented utterance. This defeats the purpose of this experiment which is to confirm if the model transforms neutral utterances properly into emotional utterances. This requires the subject to know how the neutral utterance sounds, as well as the emotional utterance. Also, because not many people are used to a synthetic voice, the quality (or lack thereof) and quirks that come with a TTS system will likely interfere with the perception of the emotion. This disqualifies the option of presenting single utterances. It also makes the option of comparing a neutral with an emotional variant more preferable, because here the test subject can first familiarise himself with the neutral variant, and thus familiarise him with the synthetic voice, and then compare it to the emotional variant of the utterance. When comparing two (or more) emotional utterances, the subject is actually judging two (or more) utterances at the same time. This is almost the same as judging two single utterances in succession. On top of that, if there are any flaws in the model that would make one fragment sound ”weird”, the subject's concentration will be influenced, which in turn will have influence on the judgement of the other utterance(s). Besides, it does not make any sense to have to judge if a happy sounding utterance sounds happier than another angry utterance sounds angry. We do not want to compare two different emotion emulations, we want to verify if each single emotion is emulated properly. Therefore the most logical option would be to present the test subject with a neutral and an emotional variant of the same utterance.

The resulting interface is shown in figure 8.1



Evaluatietest van emoties in automatisch gegenereerde spraak.

Fragment 1

Hieronder staan twee fragmenten. De eerste is een fragment van een stuk tekst zoals het TTS systeem dit standaard produceert, het tweede is geproduceerd met extra aanpassingen om het **bedroefd** te laten klinken. Klik op elk fragment om het te horen.

De tekst van dit fragment is: **Waar heb je dat vandaan gehaald?**

Dit is het standaard fragment

Dit is het aangepaste fragment

Klinkt het tweede fragment inderdaad **bedroefd**?

○ Ja
○ Nee

Zo nee, waar ligt dat aan volgens jou (wat valt op)?
Je kunt meerdere eigenschappen beoordelen.

	veel te laag	te laag	acceptabel	te hoog	veel te hoog	geen mening
Toonhoogte:	○	○	○	○	○	●
Spreeksnelheid:	○	○	○	○	○	●
Volume:	○	○	○	○	○	●

	veel te weinig	te weinig	acceptabel	te veel	veel te veel	geen mening
Toonhoogte variatie:	○	○	○	○	○	●

Overige opmerkingen:

Figure 8.1: Screenshot of the main interface

8.3 Results

The experiment was performed by 19 persons, of which two did not finish the entire experiment. The partial results of those two people have been included. The raw results are presented in tables 8.1 and 8.2. Just as with the presence test, the Krippendorff alpha value was calculated in order to determine the amount of agreement amongst the people who performed the experiment. This resulted in an alpha value of 0,256 which shows that overall, the people did not agree all that much.

Fragment	**Emotion**	**Biased**	**#Yes**	**#No**	**F0**					
					−−	−	ok	+	++	?
9	Fear	Yes	17	0	0	0	0	0	0	0
10	Sadness	Yes	11	6	0	1	3	0	0	2
26	Joy	Yes	5	12	0	0	1	6	3	2
4	Anger	Yes	14	5	0	0	2	3	0	0
7	Love	Yes	1	16	0	3	6	0	0	7
16	Surprise	Yes	9	8	0	2	2	0	0	4
14	Boredom	Yes	11	6	0	2	2	1	0	1
6	Fear	No	14	4	0	0	2	0	1	1
12	Fear	No	11	6	0	1	0	2	0	3
17	Fear	No	6	11	0	0	2	5	0	4
1	Sadness	No	9	10	1	2	3	3	0	1
2	Sadness	No	10	9	0	1	6	1	0	1
5	Sadness	No	9	10	0	2	6	0	0	2
11	Joy	No	1	16	0	0	2	11	3	0
20	Joy	No	3	14	0	0	0	5	8	1
23	Joy	No	2	15	0	0	2	3	7	3
22	Anger	No	12	5	0	0	3	1	0	1
25	Anger	No	12	5	0	1	2	2	0	0
27	Anger	No	5	12	0	1	2	4	3	2
13	Love	No	0	17	1	1	7	2	0	6
15	Love	No	1	16	1	2	1	0	0	12
24	Love	No	2	15	0	5	2	0	0	8
3	Surprise	No	7	12	0	1	6	3	0	2
8	Surprise	No	7	10	0	2	6	1	0	1
19	Surprise	No	12	5	0	0	3	1	0	1
18	Boredom	No	13	4	0	0	3	0	0	1
21	Boredom	No	12	5	0	0	3	1	0	1
28	Boredom	No	12	5	0	0	4	0	0	1

Table 8.1: Raw results of the evaluation experiment ordered by Bias and Emotion, part 1. #Yes and #No represent the amount of people that judged the emotion to be recognisable or not. The opinions on the four prosodic attributes are split between this table and table 8.2. Opinions range from way too low (−−) to way too high (++)

Next the results for the emotionally biased sentences (7 in total) were separated from the unbiased sentences (21 in total). Then the percentage of acceptance was calculated for all emotions used as the division of the number of people who accepted divided by the number of people who rejected the emotional variant of the sentence. This is shown in tables 8.3 and 8.4 for the unbiased and biased sentences respectively.

It can be fairly easily concluded from the biased sentences (and confirmed from the unbiased sentences) that the prosody changes for the emotions love and happiness do not get accepted very well. The prosody changes for the emotions anger, fear and boredom were somewhat accepted, and

Frag-	F0 Range						Speech Rate						Intensity					
ment	−−	−	ok	+	++	?	−−	−	ok	+	++	?	−−	−	ok	+	++	?
9	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
10	2	3	1	0	0	0	0	1	2	1	0	2	0	2	3	0	0	1
26	0	0	2	8	1	1	0	3	4	0	0	5	0	1	6	0	0	5
4	0	1	3	0	0	1	0	2	2	0	0	1	0	1	2	1	0	1
7	0	1	4	4	0	7	3	3	3	0	0	7	0	0	8	1	0	7
16	0	0	4	1	0	3	0	5	1	0	0	2	0	1	4	0	0	3
14	0	2	3	0	0	1	0	3	1	1	0	1	0	0	5	0	0	1
6	0	1	1	1	0	1	0	0	1	1	0	2	0	0	3	0	0	1
12	0	0	2	2	0	2	0	1	1	1	0	3	0	0	4	0	0	2
17	0	0	2	4	1	4	0	4	1	2	0	4	0	0	5	1	0	5
1	1	3	4	1	0	1	0	3	3	2	1	1	1	2	5	1	0	1
2	0	2	5	1	0	1	1	1	2	3	0	2	0	0	8	0	0	1
5	1	2	3	2	0	2	0	3	2	3	0	2	0	2	6	1	0	1
11	0	0	10	6	0	0	0	6	6	1	0	3	0	3	12	0	0	1
20	0	0	4	8	0	2	0	2	7	2	0	3	0	3	8	0	0	3
23	0	0	2	9	1	3	0	2	4	2	0	7	0	4	5	0	0	6
22	0	2	1	0	0	2	0	2	0	0	0	3	0	0	4	0	0	1
25	0	1	1	1	0	2	0	2	0	1	0	2	0	3	1	0	0	1
27	0	0	8	0	0	4	0	1	2	3	0	6	0	2	6	0	0	4
13	0	1	8	3	0	5	2	4	1	3	1	6	0	1	9	2	0	5
15	0	1	3	0	0	12	1	3	1	1	0	10	0	0	3	2	0	11
24	0	1	5	0	0	9	1	3	2	1	0	8	0	0	5	1	0	9
3	0	0	4	6	2	0	0	4	5	1	0	2	0	0	9	0	0	3
8	0	1	5	2	1	1	0	4	4	1	0	1	0	2	6	0	0	2
19	0	0	3	1	0	1	0	2	2	0	0	1	0	0	4	0	0	1
18	0	1	2	0	0	1	0	1	2	0	0	1	0	0	2	1	0	1
21	3	1	1	0	0	0	0	2	2	0	0	1	0	0	2	2	0	1
28	1	3	1	0	0	0	0	1	3	0	0	1	0	0	3	1	0	1

Table 8.2: Raw results of the evaluation experiment ordered by Bias and Emotion, part 2. The opinions on the four prosodic attributes are split between this table and table 8.1. Opinions range from way too low (−−) to way too high (++)

Emotion	sadness	surprise	anger	fear	love	happiness	boredom
Percentage	49%	49%	57%	60%	6%	12%	73%

Table 8.3: Average acceptance percentages of unbiased emotional sentences

Emotion	sadness	surprise	anger	fear	love	happiness	boredom
Percentage	65%	53%	74%	100%	6%	29%	65%

Table 8.4: Acceptance percentages of biased emotional sentences

those for the emotions sadness and surprise were fifty-fifty. The values for the biased sentences are all slightly higher (except love), which was to be expected because the semantics of the sentence pointed to the emotion, giving the test subject a nudge in the 'right' direction.

When looking at the feedback given about the nature of the non-acceptance of the sentences, as represented in tables 8.5 and 8.6 for the unbiased and biased sentences respectively and figures 8.2 to 8.5, several things become clear:

- Most obviously, the F0 and F0 Range for happiness were judged too high.
- The F0 of the emotions sadness, surprise and boredom was found acceptable.
- The intensity of all the emotions except love was found acceptable.
- Almost all people did not know what to think of the prosody of the sentences with the

emotion love, except that it was wrong.

- People generally did not agree on the speech rate of the sentences.

It should be noted that when people accepted the emotional prosody, they did not get the opportunity to provide feedback other than the general comments. This was done under the assumption that if people agreed with the prosody, there would be no need to change any parameters.

Emotion		sadness	surprise	anger	fear	love	happiness	boredom
F0	**OK**	52%	56%	32%	19%	21%	9%	71%
	Don't Know	14%	15%	14%	38%	54%	9%	21%
F0 Range	**OK**	41%	44%	45%	24%	33%	36%	29%
	Don't Know	14%	7%	36%	33%	54%	11%	7%
Speech Rate	**OK**	24%	41%	9%	14%	8%	38%	50%
	Don't Know	17%	15%	50%	43%	50%	29%	21%
Intensity	**OK**	66%	70%	50%	57%	35%	56%	50%
	Don't Know	10%	22%	27%	38%	52%	22%	21%

Table 8.5: Feedback on the prosodic attributes of the unbiased sentences.

Emotion		sadness	surprise	anger	fear	love	happiness	boredom
F0	**OK**	50%	25%	40%	N/A	38%	8%	33%
	Don't Know	33%	50%	0%	N/A	44%	17%	17%
F0 Range	**OK**	17%	50%	60%	N/A	25%	17%	50%
	Don't Know	0%	38%	20%	N/A	44%	8%	17%
Speech Rate	**OK**	33%	13%	40%	N/A	19%	33%	17%
	Don't Know	33%	25%	20%	N/A	44%	42%	17%
Intensity	**OK**	50%	50%	40%	N/A	50%	50%	83%
	Don't Know	17%	38%	20%	N/A	44%	42%	17%

Table 8.6: Feedback on the prosodic attributes of biased sentences

Lastly there are the textual comments that people entered. Because of the amount of text, the comments have been copied without altering to tables A.26 through A.53 in the appendix.

Summarising those comments, the aspect that most people had problems with was the word emphasis, which they found conflicting with what they would expect. Further comments, excluding comments about the four prosodic parameters that were changed, were about:

- more speech-rate changes within a sentence (sadness, surprise, love, happiness, boredom)
- pause durations (sadness, fear)
- alternative F0 contour suggestions (sadness, surprise, fear, love, boredom). For example, the F0 contour at the end of the surprised sentence: "Dat had je niet moeten doen." went downwards, while two people explicitly mentioned that they expected it to go up because that is what they expect from surprised utterances.
- change of articulation (sadness, love). People expected the articulation to suffer because of the experienced emotion.
- addition of non-speech elements like breathing or sighing (surprise, boredom)
- more variation in intensity, especially related to stressed words (anger)
- whispering (love)
- choice of words (love)

Stress on the wrong words, stress on the wrong syllables, too much stress, and too little stress were reported for all emotions, making it obvious that this is a big influence on how people recognise emotion. It has also been mentioned that a misplaced stress is a great influence in reducing the recognition rate, because it sounded so unnatural that it distracted people from the task, or the misplaced stress caused the perceived emotion to differ from the intended emotion. For example the sad intended sentence "Dat had je niet moeten doen." had stress on the word "niet", which

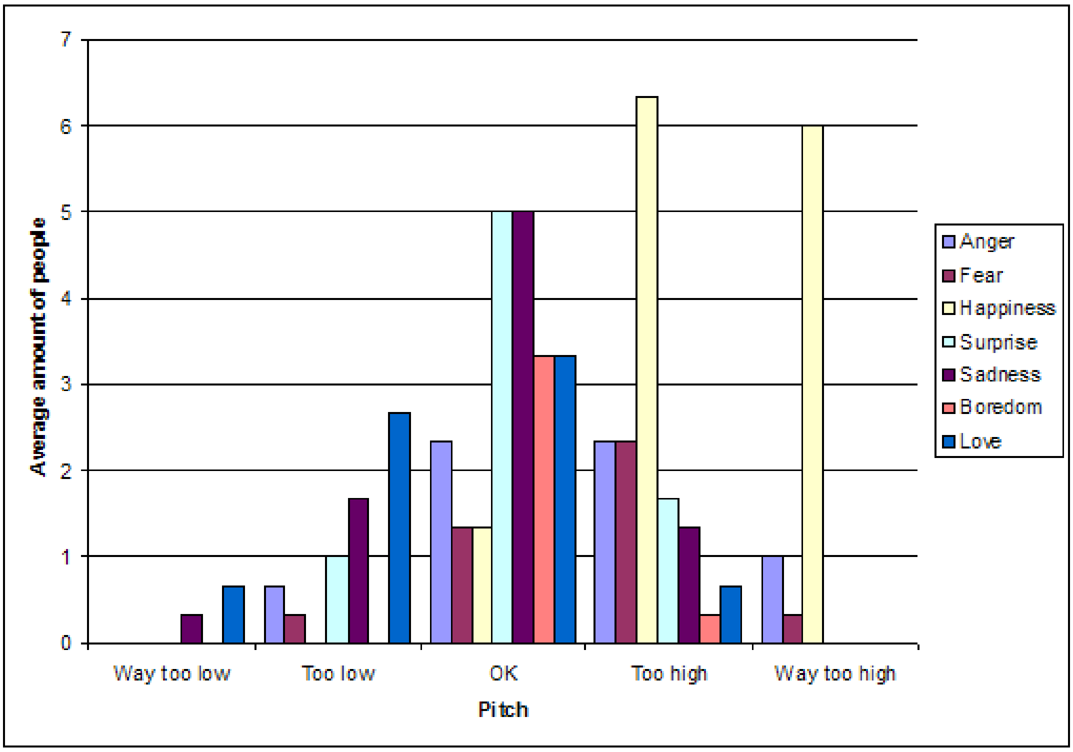


Figure 8.2: Average feedback on the pitch of the unbiased sentences

caused people to perceive it as angry instead.

8.4 Discussion

It would seem that, as with the earlier presence test, the perception of emotion is highly subjective and varies per person. This can be seen from the low Krippendorff alpha value, which shows the amount of agreement on a scale of 0 to 1, of 0.256 which indicates that most people could just as well have been randomly guessing throughout the experiment. As the various percentages of the emotion recognition rates show in table 8.7, boredom was recognised best, followed by fear and anger. The emotional variants of the sad and surprised sentences were rejected by half the people, and love and happiness were almost never accepted.

Emotion	sadness	surprise	anger	fear	love	happiness	boredom
Percentage	0.49	0.49	0.57	0.6	0.06	0.12	0.73

Table 8.7: Acceptance percentages of unbiased emotional sentences

The high acceptance rate for boredom could come from the fact that the model used was created by Mozziconacci [30] and used the Dutch language. The other models were originally used for languages other than Dutch, so it is possible that the configurations of the prosodic aspects require fine-tuning if they are applied to other languages. The low acceptance rate for love is definitely caused by the strangely low prosodic parameter changes derived by the model. This caused the neutral and emotional fragments to sound almost alike, and as such people did not find the emotional fragment sounding more loving than the neutral fragment. The low acceptance rate for happiness is caused by the extreme pitch that the model produced: +50% F0 and +100% F0 Range. This has been emphasised by the way in which the downstep contour model [2] uses

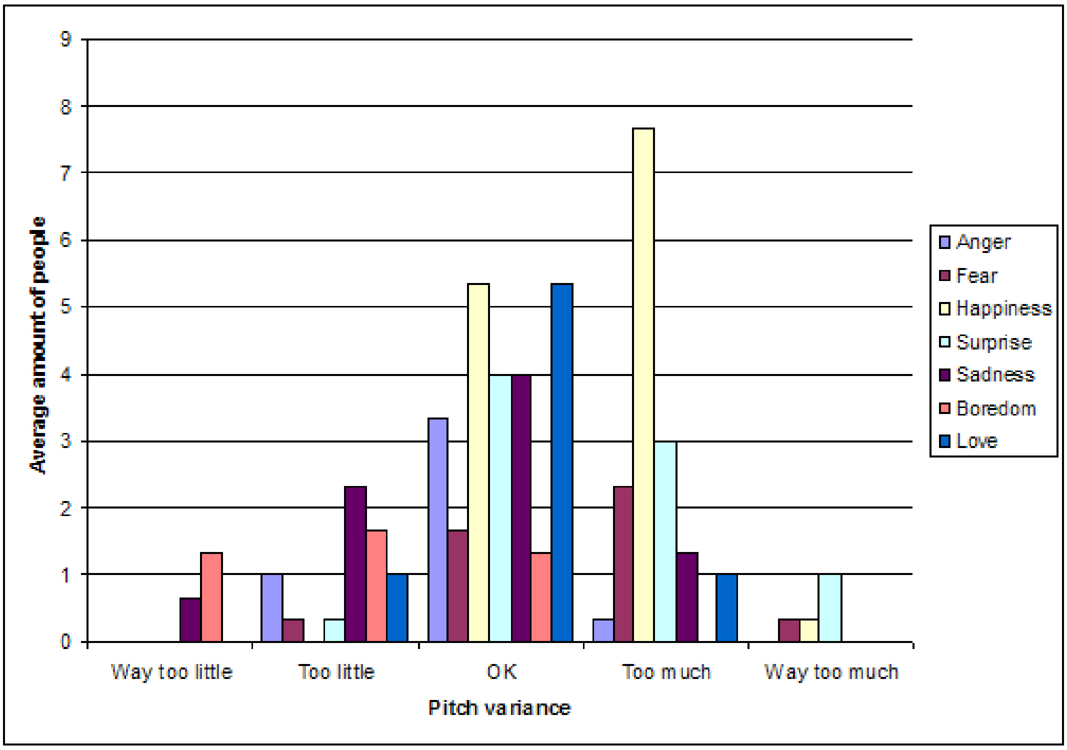


Figure 8.3: Average feedback on the pitch variance of the unbiased sentences

the F0 Range to determine the global F0 contour. This causes the F0 to rise, at the start of the sentence, by 50% (F0) + 130% (F0 Range implementation), making the total F0 extremely high.

The other rejections were mainly caused by problems people had with the accent placement. This is apparently a very important issue and will need to be addressed. The speech-rate (vowel duration and pause duration) changes that people also missed were present in the model by Schröder, but were not implemented yet because SSML does not include these changes in its specification.

8.5 Conclusion

With the exception of love and happiness, it would seem that with small improvements, the other emotions will all be recognised above chance level, given the fact that they currently are at- or above chance level already. Despite the fact that a lot of work remains to be done, these initial results look promising enough to be used in the VST system. The downstep model will have to be analysed in order to fix the F0 Range issue that caused the extremely low acceptance rate for happiness. If this is fixed, it is likely that the acceptance rate goes up to at least chance level. The speech-rate will have to be better adjustable. This implies that the current use of SSML will have to be extended to a markup language that includes methods to adjust the vowel and pause duration at will. In any case, the feedback attained from the comments opens up a lot of possibilities for future work.

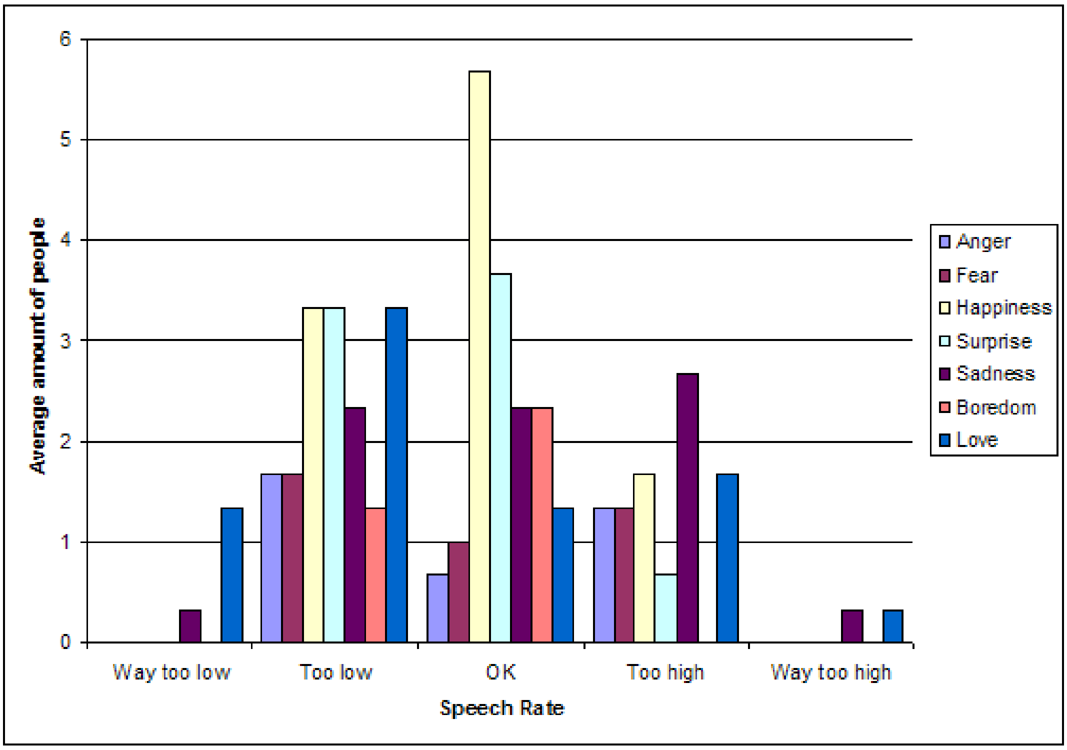


Figure 8.4: Average feedback on the speech rate of the unbiased sentences

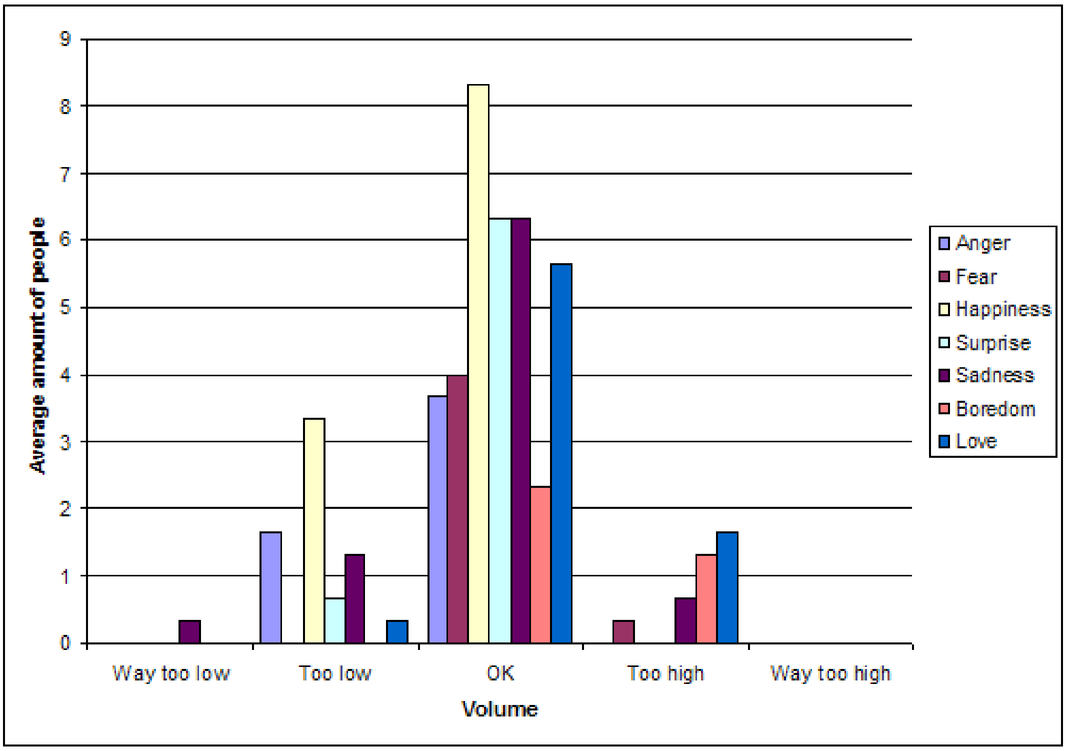


Figure 8.5: Average feedback on the volume of the unbiased sentences

Chapter 9

Conclusions and Recommendations

9.1 Summary

When this project started, the original goals were to improve the quality of the Virtual Storyteller [54] by analysing speech from a generally appreciated source of stories, and to implement the existing improvements created by Meijs [27] using an open-source Text-to-Speech engine. After a bit of literature study, the first goal was narrowed down to improving the quality of sentences uttered by in-story characters by adding emotions. Even though the VST currently does not produce character dialog, given the high degree in which such dialog is present in existing fairy-tales, it is expected that this functionality will be added to the VST sooner or later. After finding out through literature that there are multiple conflicting terminologies used for both prosody and emotions, the choice was made to use the findings of Schötz [45] for the definition of prosody, and the emotional terms determined by Parrott [37] as the method of describing emotions used in this thesis. An analysis of available Text-to-Speech engines revealed that, given the two requirements: open-source and Dutch voice support, only the Festival Speech Synthesis System[1] met these requirements. The Dutch voice support is called NeXTeNS[2].

The next step in realising the goals was to analyse speech from recorded fragments by well-known, and accepted, storytellers. This was done using fairy-tales from the Lecturama Luistersprookjes. From these fairy-tales, fragments were extracted that exhibited various emotions, as well as neutral fragments, to be used as a baseline for the neutral-to-emotional morphing model. These fragments were then presented to a group of people in a blind test in order to be labelled by the emotion these fragments were expressing. The emotions used for labelling were the primary emotions of the table by Parrott [37], as well as an 'other' emotion in which the person who participated in the test could enter other emotional terms. The agreement of the people who participated in this test was, expressed with Krippendorff's alpha-value [25] 0.284, which, on a scale of zero to one, is not much. This indicated that overall, there was not all that much agreement amongst the people. This can also be seen by the average agreement percentages, ordered by emotion in table 4.4. When comparing the perceived (and most agreed-upon) emotions with the emotions derived from the context of the story, 60% of them match. However, two of the fragments on which the agreement percentage was very high mismatched with the context. This does not invalidate the fragments for further analysis though, specifically because it is the perception of the emotion that is under investigation. Emotions were perceived at an above-chance rating for all fragments, and all characters present in those fragments had at least one fragment labelled as neutral and at least one fragment labelled as emotional, so all the fragments were used in the audio analysis.

The fragments were analysed using Praat [39] and various acoustic parameters, which according

[1] http://www.cstr.ed.ac.uk/projects/festival/

[2] http://nextens.uvt.nl/

to literature would be useful for this task, were measured. This data was then grouped per associated emotion in order to extract a neutral-to-emotion model. However, the amount of fragments turned out to be too small to compensate for the wide variety in the acoustic measurements and no model could be constructed from it. It was then decided to construct the neutral-to-emotion model using models from literature. For each of the emotions: Joy, Sadness, Anger, Fear, Surprise, and Boredom, there were existing models in literature. Also, research by Schröder [46] pointed out that using emotional terms was quite ambiguous and provided data and insight for an alternative method of describing emotions. An alternative method determines emotions on a three-axis scale: Activation, Evaluation, and Power. This method and the data were also incorporated in the model to create neutral-to-emotion transformations for emotions that are present in the VST, but of which there was not an existing model in literature. Unfortunately the data is not complete, insofar as that not all the emotional terms can yet be translated into the three-axis scale. This means that not all the emotions present in the VST are present in the model.

In order to test the model, a link between the VST and Festival must first be established. The VST can create an XML document where the text to be rendered is annotated with the desired emotions and story-telling styles. Festival does not have SSML input support, so in order to be able to signal prosodic changes for pieces of text input, partial SSML support was implemented. Also, an application was designed and implemented that translates the VST XML into SSML, which is then used by Festival to produce speech.

The method used to test the model was quite straightforward. Given the earlier "chaotic" results (when considering the Krippendorff alpha value), it was decided that the answering options of the evaluation experiment should not be as open as they were in the perception test. People participating had to agree or disagree on whether a modified fragment did indeed 'show' the target emotion. They could also provide feedback on the amount of the prosodic parameters used in the model. The Krippendorff alpha-value [25] of the results of this test turned out to be 0.256. Despite the simplification of the answering method, the overall agreement was still quite low. However, the emotion acceptance percentages in table 8.4 showed some similarity to the average emotion agreement percentages from the presence test, with the exception of the emotion "happiness" (joy). The emotions that were recognised the most were (in decreasing order): Boredom, Fear, Anger, Surprise and Sadness. The last two had an acceptance percentage of 49%. The high recognition of the emotion "boredom" may be attributed to the fact that the model for boredom was originally created and tested using the Dutch language, where the other emotional models were originally created for different languages. Compared to the neutral fragments, the changed emotional fragments were mostly noticeably different, and five emotions were accepted at a reasonable rate. This does not mean that the model is finished, especially when looking at the large amount of feedback received from the evaluation test. The model requires a lot of additional work in tweaking the parameters and further expanding the capabilities of Festival and NeXTeNS to support more alterations of prosodic parameters like manual accent placement and accent modification, speech rate changes that are more specific than merely a changed fixed speech rate, but include details like vowel- and pause duration.

Concluding, the model used is a good step in the right direction, however a lot of additional work (see the next section) is required in order to create emotional sounding speech that is accepted without doubt.

9.2 Recommendations

As is not uncommon with doing research, it raises more questions than are answered. Concerning the story-telling act, there are a few things that should be researched in order to make the VST perform more like real life storytellers:

- Fairy-tales use a lot of non-speech (but still vocal) sound effects like laughter, giggling, coughing etcetera. However, the complexity of these sound effects (having only one 'giggle' sample gets annoying very very fast) and the fact that you would require the person on who the TTS voice was based on to supply these samples, makes the inclusion of non-speech vocal sound effects quite hard. Festival has the functionality to inject digital audio samples into the synthesised speech audio-stream, so it is technically possible to use these non-speech sound effects and thus research into this is useful in order to further advance the quality of the VST.

- Storytellers emulate character voices by modifying their voice. This modification includes voice quality, and is therefore very hard to implement in todays TTS systems. More research into how storytellers change their voices, expressed in prosodic parameter changes, would be useful in order to enhance the VST with character dialog. For example, big characters (giants, elephants) have a low and hollow voice, small and fast characters usually have a high-pitched voice and talk faster. Alternatively, different TTS voices could be used to implement character voices, however changing between voices currently produces an unacceptable delay, as the voice is activated, when using the system in a real time story-telling performance.

Regarding the model, the further research into the following items is recommended to further refine the model and complete the currently incomplete data used to convert emotions to prosodic parameters:

- The current model hinges a lot on earlier research from other languages, and uses incomplete data. This data can be obtained by performing an acoustic analysis of a large corpus of Dutch emotional speech, similar to what Schröder did with the Belfast database of spontaneous emotional speech. Model B uses Schröder's method of calculating the necessary prosodic parameter changes based on the correlation between acoustic parameters and the activation, evaluation and power dimensions of the emotion dimensional framework. As such, using the same method to obtain the correlations of the acoustic parameters and emotion dimensions from a Dutch emotional speech corpus and using those correlations in Model B will quite likely sound more natural to Dutch people.
- The information used to translate emotional terms to the three-dimensional AEP scale is incomplete and based on English emotional speech. This information will need to be replaced by annotating Dutch emotional speech fragments with emotional terms as well as annotate them with the Feeltrace tool. Once all the emotion terms used in the VST can be translated into these AEP dimensions, the models that directly translate an emotional term into prosodic changes can be removed in favour of the method that uses the AEP data in combination with the analysis of Dutch emotional utterances to provide the necessary prosodic parameter changes.

The TTS system could also use some enhancement in the following areas:

- It became apparent that SSML is not expressive enough for custom syllable accent placement and accent parameter modification, as well as detailed speech rate modification. An extension to SSML must be created and implemented which can handle these detailed modifications. From the feedback of the evaluation experiment, it became apparent that accent placement is a major contributing factor towards the acceptance of emotional prosody, or a huge distraction if they are placed incorrectly. It is necessary to enhance NeXTeNS to allow for manual accent placement and accent modification which, according to comments in the source code is currently broken.
- The emotional TTS application MARY uses a voice database in which the same diphones are recorded at multiple levels of vocal effort. This aids towards changing the voice quality, something which is otherwise impossible when using TTS with a single diphone database. MARY currently supports English, German and Tibetan. This method should be investigated and implemented using Festival and Dutch voices as well.
- Various functions are used to change the frequency of the pitch, however the human perception of frequency changes is logarithmical [43] and therefore changes in pitch should be expressed in the unit st (semitones). Therefore all future work should use pitch changes in semitones instead of Hz. The conversion formula for δ Hz to st at base frequency BaseHz is:

$$st = (log_2((BaseHz + \delta Hz)/BaseHz)) * 12 \tag{9.1}$$

This also helps when using different TTS voices, which do not always have to start at the same frequency. By using semitones, the pitch increase is always perceived properly regardless of the starting frequency the change occurs at.

Bibliography

[1] C.L. Bennett. *Large Scale Evaluation of Corpus-based Synthesizers: Results and Lessons from the Blizzard Challenge 2005*
In INTERSPEECH-2005, pp. 105-108.

[2] R. van den Berg, C. Gussenhoven, and T Rietveld. *Downstep in Dutch: implications for a model.*
In G.J. Docherty and D.R. Ladd (eds.), "Papers in Laboratory Phonology II", Cambridge University Press, pp. 335-359, 1992.

[3] V. Bralé, V. Maffiolo, I. Kanellos, and T. Moudenc. *Towards an Expressive Typology in Storytelling: A Perceptive Approach*
Lecture Notes in Computer Science, volume 3784, pp. 858-865, Springer-Verlag Berlin Heidelberg, 2005.

[4] F. Burkhardt and W.F. Sendlmeier. *Verification of Acoustical Correlates of Emotional Speech using Formant-Synthesis*
In Proceedings of the ISCA Workshop on Speech and Emotion, pp. 151-156, Nothern Ireland, 2000.

[5] J.E. Cahn. *The Generation of Affect in Synthesized Speech*
Journal of the American Voice I/O Society, 8:1-19, 1990.

[6] Call of Story. *What is Storytelling?*
http://www.callofstory.org/en/storytelling/default.asp

[7] N. Campbell and T. Marumoto. *Automatic labelling of voice-quality in speech databases for synthesis*
In Proceedings of the 6th Internatonal Conference on Spoken Language Processing, Beijing, China, 2000.

[8] G.A. Chesin. *Storytelling and Storyreading*
In Peabody Journal of Education: Vol. 43, No. 4, pp. 212- 214, 1966.

[9] R. Cowie, E. Douglas-Cowie, B. Apolloni, J. Taylor, A. Romano, and W. Fellenz. *What a neural net needs to know about emotion words*
In N. Mastorakis, editor, Computational Intelligence and Applications, pages 109-114. World Scientific & Engineering Society Press, 1999.

[10] R. Cowie, E. Douglas-Cowie, S. Savvidou, E. McMahon, M. Sawey, and M. Schröder. *'Feeltrace': An Instrument For Recording Perceived Emotion in Real Time*
In Proc. ISCA ITRW on Speech and Emotion: Developing a Conceptual Framework, Belfast, pp 19-24, 2000.

[11] L. Deng, D. Yu, and A. Acero. *A Quantitative Model for Formant Dynamics and Contextually Assimilated Reduction in Fluent Speech*
In INTERSPEECH, pg. 981-984, 2004.

[12] E. Douglas-Cowie, R. Cowie, and M. Schröder. *A new emotion database: considerations, sources and scope.*
In Proceedings of the ISCA Workshop on Speech and Emotion, pp. 39-44, Northern Ireland, 2000.

[13] T. Dutoit. *High Quality Text-to-Speech Synthesis: A Comparison of Four Candidate Algorithms*
In Proceedings of ICASSP pp. 565-568, 1994.

[14] E. Eide, R. Bakis, W. Hamza, and J. Pitrelli. *Multilayered Extensions to the Speech Synthesis Markup Language for Describing Expressiveness*
in EUROSPEECH-2003, pp. 1645-1648.

[15] P. Ekman. *Emotion in the human face, second edition.*
Cambridge University Press, New York, 1982.

[16] National Storytelling Association. *What is Storytelling?*
http://www.eldrbarry.net/roos/st_defn.htm, 1997.

[17] S. Faas. *Virtual Storyteller: An approach to computational story telling*
M.Sc. Thesis, University of Twente, Enschede. The Netherlands. June 2002.

[18] Author Unknown. *Storytelling: Definition & Purpose*
http://falcon.jmu.edu/r̃amseyil/storydefinition.htm

[19] N.H. Frijda. *The Emotions*
Cambridge University Press, Cambridge, England, 1986.

[20] C. Gobl and A. Ní Chasaide. *Testing affective correlates of voice quality through analysis and resynthesis*
In Proceedings of the ISCA Workshop on Speech and Emotion, pp 178-183, Northern Ireland, 2000.

[21] B. Heuft, T. Portele, and M. Rauth. *Emotions in time domain synthesis*
In Proceedings of the 4th International Conference of Spoken Language Processing, Philadelphia, USA, 1996.

[22] I. Iriondo, R. Guaus, A. Rodríguez, P. Lázaro, N. Montoya, J.M. Blanco, D. Bernadas, J.M. Oliver, D. Tena, and L. Longhi. *Validation of an acoustical modelling of emotional expression in Spanish using speech synthesis techniques*
in Proceedings of the ISCA Workshop on Speech and Emotion, pp. 161-166, Northern Ireland, 2000.

[23] C.E. Izard. *Human emotions*
Plenum Press, New York, 1977.

[24] Sun Microsystems, Inc. *JSpeech Markup Language*
http://java.sun.com/products/java-media/speech/forDevelopers/JSML/index.html, version 0.6, 1999.

[25] K. Krippendorff. *Content Analysis, an introduction to Its Methodology.*
Thousand Oaks, CA: Sage Publications, 1980.

[26] S. Lemmetty. *Review of Speech Synthesis Technology*
M.Sc. Thesis, Helsinki University of Technology, Department of Electrical and Communications Engineering, 1999.

[27] K. Meijs. *Generating natural narrative speech for the Virtual Storyteller*
M.Sc. Thesis, Human Media Interaction Group. Department of Electrical Engineering, Mathematics and Computer Science. University of Twente, Enschede. The Netherlands. March 2004.

[28] J.M. Montero, J. Gutiérrez-Arriola, S. Palazuelos, E. Enríquez, S. Aguilera, and J.M. Pardo. *Emotional Speech Synthesis: From Speech Databse to TTS*
In Proceedings of the 5th International Conference on Spoken Language Processing, volume 3, pp. 923-926, Sydney, Australia, 1998.

[29] J.M. Montero, J. Gutiérrez-Arriola, J. Colás, E. Enríquez, and J.M. Pardo. *Analysis and modelling of emotional speech in Spanish*
In Proceedings of the 14th International Conference of Phonetic Sciences, pp. 957-960, San Francisco, USA, 1999.

[30] S.J.L. Mozziconacci. *Speech Variability and Emotion: Produciton and Perception.*
PhD thesis, Technical University Eindhoven, 1998.

[31] S.J.L. Mozziconacci. *Modeling Emotion and Attitude in Speech by Means of Perceptually Based Parameter Values*
in User Modelling and User-Adapted Interaction 11, pp 297-326, 2001.

[32] S.J.L. Mozziconacci and D.J. Hermes. *Role of intonation patterns in conveying emotion in speech*
In Proceedings of the 14th International Conference of Phonetic Sciences, pp. 2001-2004, San Francisco, USA, 1999.

[33] I.R. Murray, M.D. Edgington, D. Campion, and J. Lynn. *Rule-based emotion synthesis using concatenated speech*
In Proceedings of the ISCA Workshop on Speech and Emotion, pages 173-177, Northern Ireland, 2000.

[34] I.R. Murray and J.L. Arnott. *Implementation and testing of a system for producing emotion-by-rule in synthetic speech*
Speech Communication, 16:369-390, 1995.

[35] E. Navas, I. Hernáez, and I. Luengo. *An Objective and Subjective Study of the Role of Semantics and Prosodic Features in Building Corpora for Emotional TTS*
IEEE Transactions on Audio, Speech, and Language Processing, Vol. 14, No. 4, pp 1117-1127, July 2006.

[36] A. Ortony, G.L. Clore, and A. Collins. *The Cognitive Structure of Emotion*
Cambridge University Press, Cambridge, UK, 1988.

[37] W. Parrott. *Emotions in Social Psychology*
Psychology Press, Philadelphia, 2001.

[38] R. Plutchik. *Emotion: a psychoevolutionary synthesis.*
Harper & Row, New York, 1980.

[39] P. Boersma and D. Weenink. *Praat: doing phonetics by computer*
Version 4.5.02 [Computer program], http://www.praat.org/, 2007.

[40] E. Rank. *Erzeugung emotional gefärbter Sprache mit dem VieCtoS-Synthesizer.*
Technical Report 99-01, ÖfAI, http://www.ofai.at/cgi-bin/tr-online?number+99-01, 1999.

[41] E. Rank and H. Pirker. *Generating Emotional Speech with a Concatenative Synthesizer*
In Proceedings of the International Conference of Speech and Language Processing, Vol. 3, pp. 671-674, Sydney, Australia, 1998.

[42] S.H.J. Rensen. *De virtuele verhalenverteller: Agent-gebaseerde generatie van interessante plots*
M.Sc. thesis, Faculty of Computer Science, University of Twente, Enschede, March 2004.

[43] A.C.M. Rietveld and V.J. van Heuven. *Algemene Fonetiek*
Coutinho, Bussum, 1997.

[44] R. Sproat, A. Hunt, M. Ostendorf, P. Taylor, A. Black, K. Lenzo, and M. Edgington. *SABLE: A STANDARD FOR TTS MARKUP*
Bell Labs - Lucent Technologies, Sun Microsystems, Inc. Boston University, CSTR - University of Edingburgh, Carnegie-Mellon University, BT Labs.

[45] S. Schötz. *Prosody in Relation to Paralinguistic Phonetics - Earlier and Recent Definitions, Distinctions and Discussions*
Term paper, Department of Linguistics and Phonetics, Lund University, 2003.

[46] M. Schröder. *Speech and Emotion Research: An overview of Research Frameworks and a Dimensional Approach to Emotional Speech Synthesis*
PhD thesis, Vol. 7 of Phonus, Research Report of the Institute of Phonetics, Saarland University, Germany, 2004.

[47] N. Slabbers. *Narration for Virtual Storytelling*
M.Sc. thesis, Human Media Interaction Group, Department of Electrical Engineering, Mathematics and Computer Science, University of Twente, Enschede, The Netherlands, March 2006.

[48] C. Sobin and M. Alpert. *Emotion in Speech: The Acoustic Attributes of Fear, Anger, Sadness, and Joy*
In Journal of Psycholinguistic Research, Vol. 28, No. 4, pp. 347-365, 1999.

[49] W3C Consortium. *Speech Synthesis Markup Language (SSML) Version 1.0*
http://www.w3.org/TR/speech-synthesis/ September 2004.

[50] Various Authors. *The Storytelling FAQ*
http://www.timsheppard.co.uk/story/faq.html, 2003.

[51] B.W. Sturm. *The 'Storylistening' Trance Experience*
Journal of Americal Folklore, 113, pp. 287-304, 2000.

[52] Sun Microsystems. *API Markup Language Specification* Version 0.6
http://java.sun.com/products/java-media/speech/forDevelopers/JSML/index.html October 1999.

[53] J. Tao, Y. Kang, and A. Li. *Prosody Conversion From Neutral Speech to Emotional Speech*
IEEE Transactions on Audio, Speech, and Language Processing, Vol 14, No. 4, pp 1145-1154, July 2006.

[54] M. Theune, S. Faas, A. Nijholt, and D. Heylen. *The Virtual Storyteller*
In ACM SIGGROUP Bulletin, Volume 23, Issue 2, ACM Press, pages 20-21, 2002.

Appendix A

Tables

A.1 Presence test

Nr	Character	Sentence
1	Tante	Die onbekende prinses is toch verdwenen.
2	Dot	Ik vind het anders helemaal niet grappig
3	Tante	Dat gebeurt misschien nog wel.
4	Magda	En daarin staat dat we drie wensen mogen doen
5	Man	Wat is er met jou aan de hand?
6	Dot	Ja, ik ben de weg kwijtgeraakt.
7	Man	Wat ben je toch lui, Magda.
8	Dot	Dank u wel voor uw hulp.
9	Man	Nee, nee, het is jouw schuld niet lieverd.
10	Man	Waarom hebben we ook zo'n ruzie gemaakt over die wensen?
11	Magda	Er is een brief van de feeën gekomen.
12	Man	Wat? Is er geen eten?
13	Magda	Luister wat ik heb bedacht.
14	Dot	Ik heb nog nooit een vogelbekdier gezien.
15	Tante	Wat jij moet aantrekken?
16	Dot	Ik u ook.
17	Dot	Wie zijn dat?
18	Dot	Maar dat is de baby van mijn kangoeroe.
19	Man	We moeten heel goed nadenken, Magda.
20	Dot	Maar iemand moet toch weten waar mijn huis is?
21	Dot	Oh lieve kangoeroe.
22	Magda	Ik heb al een lijstje gemaakt.
23	Dot	Ik heb mensen nog nooit zo raar zien doen.
24	Dot	Laat mij toch hier.
25	Man	Ik zou best een paar worstjes lusten.
26	Tante	Ik zal het eerst proberen.
27	Dot	Kangoeroe, laat mij maar uitstappen.
28	Magda	Kom binnen en doe de deur dicht
29	Magda	Dat wens ik ook.
30	Magda	Wat ben je toch dom.

Table A.1: Sentences presented in the presence test, in order of appearance

Nr	None	Happiness	Sadness	Fear	Love	Anger	Surprise	Other	Character
6	9	3	8	8	2	0	0	2	Dot
18	1	0	1	0	0	20	3	7	Dot
2	0	0	26	0	1	3	0	2	Dot
14	13	4	0	0	0	0	14	1	Dot
20	2	0	14	15	0	0	0	1	Dot
17	5	0	1	13	1	0	6	6	Dot
23	7	1	0	3	0	0	17	4	Dot
27	5	0	1	22	2	0	0	2	Dot
24	2	1	17	11	0	1	0	0	Dot
21	2	2	18	1	5	0	0	4	Dot
8	2	27	1	0	0	0	0	2	Dot
16	15	0	6	3	6	1	0	1	Dot
28	9	0	1	11	2	0	3	6	Magda
29	13	1	11	1	1	4	0	1	Magda
11	8	2	0	0	0	0	14	8	Magda
4	2	17	0	1	0	0	7	5	Magda
22	19	5	1	0	2	0	2	3	Magda
13	18	7	0	1	0	0	0	6	Magda
30	3	0	4	1	2	20	0	2	Magda
5	7	0	2	0	0	5	10	8	Frederik
25	15	9	0	1	0	2	0	5	Frederik
19	22	0	1	2	0	1	0	6	Frederik
12	0	0	0	0	1	9	18	4	Frederik
7	2	0	1	0	0	27	0	2	Frederik
10	1	1	16	0	0	7	0	7	Frederik
9	19	0	0	0	6	1	1	5	Frederik
15	4	0	0	1	0	12	7	8	Tante
3	14	0	0	0	1	5	2	10	Tante
1	13	0	4	3	0	6	0	6	Tante
26	16	5	1	0	0	1	1	8	Tante

Table A.2: Raw results of the presence test, ordered by character. The numbers represent the amount of test subjects that chose each emotion per question.

Character	Question	Question text
Text entered by test subjects as 'Other' emotion		
Dot	6	Ja, ik ben de weg kwijtgeraakt.
Schaamte		
Dot	18	Maar dat is de baby van mijn kangoeroe.
'Het is oneerlijk', 2x Jaloezie, verontwaardiging, verongelijkt, irritatie		
Dot	2	Ik vind het anders helemaal niet grappig
Verontwaardiging, irritatie		
Dot	14	Ik heb nog nooit een vogelbekdier gezien.
Verbazing		
Dot	20	Maar iemand moet toch weten waar mijn huis is?
Wanhoop		
Dot	17	Wie zijn dat?
2x Nieuwschierigheid, verlegen/angst, anticipatie		
Dot	23	Ik heb mensen nog nooit zo raar zien doen.
Opgewonden, 3x verbazing		
Dot	27	kangoeroe, laat mij maar uitstappen.
'Zielig kindje'		

Continued on next page

Table A.3 – continued from previous page

Character	Question	Question text
Text entered by test subjects as ’Other’ emotion		
Dot	24	Laat mij toch hier.
Dot	21	Oh lieve kangoeroe.
’Blij en droef’, opluchting, ’mix tussen blij, droef en liefde’		
Dot	8	Dank u wel voor uw hulp.
Opgelucht, dankbaarheid		
Dot	16	Ik u ook.
Magda	28	Kom binnen en doe de deur dicht
Samenzweerderig, voorzichtig, spanning, 2x geheimzinnig		
Magda	29	Dat wens ik ook.
Hoop		
Magda	11	Er is een brief van de feeën gekomen.
Samenzweerderig, 2x spanning, enthousiasme, opgewekt, 2x opgewonden, anticipatie		
Magda	4	En daarin staat dat we drie wensen mogen doen
2x Opgewonden, verrukking/enthousiasme, zekerheid		
Magda	22	Ik heb al een lijstje gemaakt.
hoop, enthousiasme/verassing, kalmte		
Magda	13	Luister wat ik heb bedacht.
Opgetogen, trots, opgewonden, vrolijkheid, zekerheid		
Magda	30	Wat ben je toch dom.
Teleurstelling		
Man	5	Wat is er met jou aan de hand?
Nors, chagrijnig, bazig, desinteresse, verbazing		
Man	25	Ik zou best een paar worstjes lusten.
Gedrevenheid, hongerig, knorrig, hoop (+anticipatie)		
Man	19	We moeten heel goed nadenken, Magda.
2x bedachtzaam, 2x ernst		
Man	12	Wat? Is er geen eten?
2x Verbazing		
Man	7	Wat ben je toch lui, Magda.
Teleurstelling (negatief richting boos), Irritatie		
Man	10	Waarom hebben we ook zo’n ruzie gemaakt over die wensen?
3x Spijt, teleurstelling, berouw		
Man	9	Nee, nee, het is jouw schuld niet lieverd.
2x Troost, 2x geruststellend		
Tante	15	Wat jij moet aantrekken?
3x Verontwaardiging, bazig, irritatie of verbazing, 2x verbazing		
Tante	3	Dat gebeurt misschien nog wel.
2x Hoop, 6x Irritatie, Kalmte		
Tante	1	Die onbekende princes is toch verdwenen.
Onbezorgdheid, onverschilligheid, arrogantie, apathy, bevestiging		
Tante	26	Ik zal het eerst proberen.
2x Vastberadenheid, ijverzucht, bazig, eigenzinnig, zelfverzekerd, minachting		

Table A.3: This table contains the text of each fragment in the presence test, as well as the text entered in the ’Other’ emotion field. The entries are ordened by character.

Question	Emotion	Agreement Before	Agreement After	Change
6	None	0.28125	0.28125	0
18	Anger	0.625	0.8125	0.1875
2	Sadness	0.8125	0.8125	0
14	Surprise	0.4375	0.5	0.0625
20	Fear	0.46875	0.5	0.03125
17	Fear	0.40625	0.40625	0
23	Surprise	0.53125	0.59375	0.0625
27	Fear	0.6875	0.6875	0
24	Sadness	0.53125	0.53125	0
21	Sadness	0.5625	0.59375	0.03125
8	Happiness	0.84375	0.84375	0
16	None	0.46875	0.46875	0
28	Fear	0.34375	0.34375	0
29	None	0.40625	0.40625	0
11	Surprise	0.4375	0.5	0.0625
4	Happiness	0.53125	0.625	0.09375
22	None	0.59375	0.59375	0
13	None	0.5625	0.5625	0
30	Anger	0.625	0.65625	0.03125
5	Surprise	0.3125	0.28125	-0.03125
25	None	0.46875	0.46875	0
19	None	0.6875	0.6875	0
12	Surprise	0.5625	0.5625	0
7	Anger	0.84375	0.90625	0.0625
10	Sadness	0.5	0.5	0
9	None	0.59375	0.5625	-0.03125
15	Anger	0.375	0.40625	0.03125
3	None	0.4375	0.4375	0
1	None	0.40625	0.40625	0
26	None	0.5	0.5	0

Table A.4: Counting ’other’ emotions towards their primary emotion. The agreement values represent the fraction of people that have chosen this specific answer with respect to the total amount of people that have answered that question. The ’Agreement Before’ value represents the agreement before- and ’Agreement After’ represents the agreement after counting various ’Other’ emotion answers towards a primary emotion. The list of ’Other’ emotion descriptions and to which primary emotion they were counted towards are listed in table A.5

Character	Sentence	Context	Emotion
Dot	6	(Dot is verdwaald, begint te huilen en komt kangoeroe tegen die haar bessen voert. Hierdoor kan Dot de dieren verstaan.) Toen hoorde ze een stem die duidelijker klonk dan de anderen. Het was de kangoeroe die tegen haar zei ”Ik ben heel verdrietig, want ik ben mijn baby-kangoeroe kwijtgeraakt. Ben jij ook iets kwijtgeraakt?” -”Ja”, fluisterde Dot, *”Ik ben de weg kwijtgeraakt”*	Sadness

Continued on next page

Table A.6 – continued from previous page

Character	Zin	Context	Emotie
Dot	2	(Dot is in slaap gevallen en wordt wakker met een enorme slang op haar buik.) Dot voelde haar hart bonzen. Opeens hoorde ze buiten heel hard lachen en roepen "Hihi, wat een grap. Maar je hoeft niet bang te zijn, als je stil blijft liggen gebeurt er niets. Ik zal de slang doodmaken, hahaha, wat een grap." Dot zag buiten de grot een vogel op de onderste tak van een boom zitten. Het was een reuze-ijsvogel, maar omdat'ie steeds zo lachte noemde Dot hem een lachvogel. *"Ik vind het anders helemaal niet grappig"*, zei Dot.	Anger
Dot	14	Onderweg zei Dot tegen de kangoeroe: *"Ik heb nog nooit een vogelbekdier gezien.* Hoe zit'ie eruit?"	None
Dot	20	Dot vertelde het vogelbekdier dat ze de weg naar huis was kwijtgeraakt. Maar het was alsof het vogelbekdier niet naar haar luisterde. Hij bleef maar mopperen over de mensen. *"Maar iemand moet toch weten waar mijn huis is?"*, riep Dot wanhopig.	Fear
Dot	17	Midden in de nacht werd Dot wakker. De maan stond hoog aan de hemel. Ze zag de kangoeroe angstig heen en weer lopen. In de verte klonk het geroffel van trommels. *"Wie zijn dat?"* , vroeg Dot. "Dat zijn jagers", zei de kangoeroe, "We moeten vluchten."	None
Dot	23	"Maar ze zullen ons toch geen kwaad doen?", zei Dot."Bovendien zou ik de jagers best willen zien." "Kom dan maar mee, maar je moet wel heel stil zijn.", zei de kangoeroe. Dot klom in de buidel en de kangoeroe verdween tussen de struiken. Ze verstopten zich achter een rots. Dot gluurde over de rand van de buidel. Ze zag een groot vuur en daar omheen dansten mannen die zich beschilderd hadden met rode en witte strepen. Ze hadden lange speren in hun hand. "Ik ben bang", fluisterde Dot. *"Ik heb mensen nog nooit zo raar zien doen."*	Fear
Dot	27	Opeens begon een hond van de jagers te blaffen. De kangoeroe schrok en sprong weg. De jagers zagen haar en renden achter haar aan. Ze konden de kangoeroe heel goed zien omdat de maan helder scheen. De honden van de jagers blaften en kwamen steeds dichterbij. De kangoeroe begon moe te worden. *"kangoeroe", riep Dot, "Laat mij maar uitstappen,* zonder mij kun je veel harder lopen."	None

Continued on next page

Table A.6 – continued from previous page

Character	Zin	Context	Emotie
Dot	24	De kangoeroe wist dat er nog maar één kans was om te ontsnappen. Ze moest naar de rots aan de overkant springen. Ze pakte Dot op, en zette haar weer in de buidel. *"Laat mij toch hier."*, riep Dot.	Sadness
Dot	21	Maar de kangoeroe nam een aanloop en sprong over het ravijn. Dot hield haar adem in. De kangoeroe landde met haar voorpoten op de rand van de rots aan de andere kant. Ze probeerde zich omhoog te trekken, maar dat lukte niet. Langzaam zakte ze langs de steile rotswand omlaag. De kangoeroe probeerde met haar achterpoten ergens steun te vinden. Gelukkig lukte dat. Voorzichtig klom ze naar de rand van de rots en viel toen uitgeput voorover. Dot klom uit de buidel en sloeg haar armen om de hals van de kangoeroe. *"Oh lieve kangoeroe."*, huilde ze.	Sadness
Dot	8	(De roerdomp geeft advies om de kangoeroe weer bij te brengen. Dot doet dit en het werkt.) *"Dank u wel voor uw hulp."*, zei Dot tegen de roerdomp.	Happiness
Dot	16	Opgewonden zei ze tegen Dot "Ik heb het kwikstaartje gevonden en hij kan je de weg naar huis wijzen.". Ze nam Dot mee naar het kwikstaartje. "Dag Dot", zei het vogeltje. "Een heleboel mensen zijn je aan het zoeken, en ze zijn allemaal heel verdrietig. Het is nu te donker om te vertrekken, maar morgenvroeg zal ik je thuisbrengen." Dot en de kangoeroe bleven die avond nog lang praten. "Ik zal je missen", zei de kangoeroe. *"Ik u ook."*, zei Dot	Sadness
Dot	18	(Dot komt thuis; de kangoeroe loopt mee.) Dot en haar ouders liepen naar de boerderij, en de kangoeroe liep achter hen aan. De deur van de boerderij stond open, en daar kwam een baby-kangoeroe naar buiten huppelen die meteen in de buidel van de kangoeroe sprong. "Kijk nu toch eens", zei de moeder van Dot. "Huppel is in de buidel van de grote kangoeroe gesprongen." "Huppel?", vroeg Dot. "Vader heeft vorige week in een struik een baby kangoeroe gevonden", zei moeder. *"Maar dat is de baby van mijn kangoeroe"*, riep Dot. "En wat zijn ze blij dat ze elkaar weer gevonden hebben"	Surprise

Continued on next page

Table A.6 – continued from previous page

Character	Zin	Context	Emotie
Frederik	5	Op een avond kwam Frederik thuis. Hij was moe en had een boze bui. Z'n vrouw Magda zat in de keuken. Ze keek'm heel vreemd aan. Op haar schoot lag een verkreukelde brief. "Wat is er met jou aan de hand?", bromde Frederik	Anger
Magda	28	"Kom binnen en doe de deur dicht", zei Magda geheimzinnig.	Fear
Magda	11	Er is een brief van de feeën gekomen.	A bit of residual fear (see above) from the suspense, and enthusiasm/happiness (good news)
Magda	4	En daarin staat dat we drie wensen mogen doen	Idem
Frederik	19	Frederik pakte de brief en las hem. "We moeten heel goed nadenken, Magda.", zei hij. "We kunnen rijk en beroemd worden, maar dan moeten we wel de juiste dingen wensen."	None
Magda	22	"Ik heb al een lijstje gemaakt.", zei Magda, en sprong op.	Enthusiast, Self-assured, No clear emotion
Magda	13	"Luister wat ik heb bedacht."	Idem
Frederik	12	[lijst van wensen].."Oh help, ik heb er helemaal niet aan gedacht om te koken." -"Wat?!", riep Frederik, "Is er geen eten? Hoe kan ik nu een wens bedenken met een lege maag?"	Quite unpleasantly surprised, Anger and Surprise (but Anger is more prominent)
Frederik	7	"Wat ben je toch lui, Magda."	Anger
Frederik	25	Ik zou best een paar worstjes lusten.	None
Magda	30	[plop, wens vervuld] "Je hebt een wens verknoeid!", riep Magda kwaad, "Wat ben je toch dom. Oh, wat maak je me boos. Ik..ik..zou willen dat die worstjes aan die rare neus van je hingen!"	Anger
Frederik	10	"Oh, wat was ik gelukkig toen ik nog een gewone neus had. Waarom hebben we ook zo'n ruzie gemaakt over die wensen?"	Compassionate, Regret → Sadness
Frederik	9	"Je hebt gelijk Frederik, het spijt me erg", zei Magda. "Nee, nee, het was jouw schuld niet lieverd.", zei Frederik	Love
Magda	29	"Ik wilde maar dat de feeën hun wensen zelf gehouden hadden en dat hier alles hetzelfde gebleven was." -"Dat wens ik ook.", zei Magda	None

Continued on next page

Table A.6 – continued from previous page

Character	Zin	Context	Emotie
Tante	15	(Er zijn uitnodigingen voor het bal van de prins. Assepoester vraagt wat zij aan zal gaan trekken.) Drie verbaasde gezichten staarden haar aan. ”Jij? Wat jij moet aantrekken?”, zei haar stiefmoeder, ”Je denkt toch niet dat jij naar het bal gaat?!”	Surprised, Angry, malicious delight → Anger
Tante	3	(Het bal is afgelopen; iedereen is weer thuis.) ”Het is allemaal de schuld van Assepoester”, zeurde Bertha, ”Als ze mijn jurk netter had gestreken was de prins zeker verliefd op mij geworden.” -”En als ze mijn pruik beter had geborsteld, was ik vast zijn vrouw geworden.”, snauwde Truida. -”Dat gebeurt misschien nog wel”, zei hun moeder, ”Die onbekende prinses is toch verdwenen”.	Sounding like ’quiet, quiet (shut up!) it’ll be all right’. No prominent emotion
Tante	1	”Die onbekende prinses is toch verdwenen”	Idem (see entry above)
Tante	26	Er werd op de deur geklopt. Buiten stond de dienaar van de prins. Hij droeg het glazen muiltje op een roodfluwelen kussen. Bertha trok hem naar binnen. ”Geef hier”, riep ze. ”Nee, ik eerst”, schreeuwde Truida. ”Uit de weg”, zei hun moeder, ”Ik zal het eerst proberen”.	A little bit angry, but mostly ’bossy’, conceited, wanting to execute her plan. No prominent emotion.

Table A.6: Contextually determined emotions for all fragments in chronological order

Amount	**Other description**	**Emotion**
9	Irritatie	Anger
2	Opgewonden	Happiness
1	Verrukking/Enthousiasme	Happiness
2	Boos, Nors	Anger
1	Chagrijn	Anger
1	Schaamte	Sadness
2	Teleurstelling (wel negatief richting boos)	Anger
2	Troost	Love
2	Geruststelling	Love
1	Anticipatie	Surprise
1	Spanning	Surprise
2	Verontwaardigd	Anger
3	Verbazing	Surprise
1	Het is oneerlijk!	Anger
2	Jaloezie	Anger
1	Wanhoop	Fear

Table A.5: 'Other' emotion modifications. Which emotion description entered in the 'Other' emotion field of the presence test is linked with which primary emotion is shown here. The frequency of occurrance is also shown.

A.2 The Model

Emotion	**Fear**	**Sadness**	**Surprise**	**Anger**	**Happiness**
Max F0	1.222541571	1.176908446	1.237693444	1.446686958	1.368386684
Mean F0	1.22096678	1.23665635	0.986873708	1.295474026	1.169113631
F0 Range	1.227974598	1.068087732	1.624904677	1.769815201	1.798990234
StdDevF0	1.144609034	1.012592061	1.594976358	1.487377441	1.612019038
Max dF0	1.764232076	1.578677128	2.513194704	3.099383509	2.698161157
Mean dF0	0.149527059	0.568871714	-0.03690214	0.502971066	-0.174601861
dF0 Range	1.262625257	1.198860051	1.52000323	1.721444638	1.583566202
StdDevdF0	1.336404415	1.117496082	1.261316472	1.241494057	1.739518713
Max Int	1.013216473	1.042599479	1.012666159	1.049024599	1.038589472
Mean Int	1.01659747	1.020614328	0.980106942	1.040509672	1.042852761
Int Range	1.125848395	1.240066621	1.245496503	1.380150725	1.064411834
StdDev Int	1.08943575	1.083128561	0.929914123	1.186451268	1.051758967
Max dInt	1.670342008	1.373827918	1.688307687	1.759963975	1.274011138
Mean dInt	0.489232713	0.645709265	0.638872538	1.049197917	0.752841381
dInt Range	1.423211049	1.363003352	1.736252097	1.578373138	1.465878257
StdDev dInt	1.387627348	1.302425219	1.297319177	1.430406868	1.171174659
Jitter (local)	1.00052789	0.891606546	0.880344888	0.605666021	0.724265353
Shimmer (local)	0.826593234	0.804091267	0.867820614	0.597403619	0.636428009
Fraction of locally unvoiced frames	0.654995227	0.424314462	0.685804586	0.487935056	0.272272526
Mean harmonics-to-noise ratio (dB)	1.151753104	1.292366691	1.170927684	1.772096421	1.372169467
Speech Rate	1.193327469	1.283287789	1.079747999	1.205114685	1.123126667
End Frequency (Hz)	1.480549012	1.318385599	1.447170816	1.129407833	1.56099078

Table A.7: Average multiplication factor for acoustic attributes between neutral and emotional for character Dot

Emotion	**Fear**	**Surprise**	**Anger**	**Happiness**
Max F0	0.890946353	0.901228283	1.348455543	1.135880073
Mean F0	1.07307999	0.901691465	1.702473667	0.989371087
F0 Range	0.61126367	0.956089217	1.080763621	1.37465485
StdDevF0	0.452049268	0.854298815	1.080507285	1.436863831
Max dF0	1.305834232	1.541143376	1.224249931	2.132524287
Mean dF0	0.398043712	1.025169677	0.858002797	0.449972053
dF0 Range	1.117956755	1.208712793	1.086490626	1.436801627
StdDevdF0	1.031215753	0.961134934	1.295703086	1.113682286
Max Int	0.964502855	0.938664771	1.065208966	1.015050033
Mean Int	0.996704345	0.940415721	1.058104753	0.987905108
Int Range	0.764244353	0.737648958	1.163385399	0.947564396
StdDev Int	0.738180981	0.519817765	1.209601489	0.995071354
Max dInt	0.397925232	0.728139234	1.092853739	1.07013404
Mean dInt	0.867955482	0.680352088	0.995869964	0.747961285
dInt Range	0.610153628	0.700741642	0.884988776	0.888312683
StdDev dInt	0.559125593	0.618342965	0.913035978	0.759933716
Jitter (local)	0.864892793	0.905450788	0.495995867	0.77680186
Shimmer (local)	0.980676153	1.050838822	0.708759416	0.837948979
Fraction of locally unvoiced frames	1.167701738	0.829682814	0.714157359	0.91677659
Mean harmonics-to-noise ratio (dB)	1.303300689	0.984403337	1.80137831	1.384983678
Speech Rate	0.827697782	1.022271831	0.969561539	0.799737037
End Frequency (Hz)	1.533635516	0.795571655	2.693629421	0.795947031

Table A.8: Average multiplication factor for acoustic attributes between neutral and emotional for character Magda

Emotion	**Sadness**	**Surprise**	**Anger**
Max F0	1.586019623	1.4748666	0.91095222
Mean F0	1.535743574	1.325401347	0.974543103
F0 Range	1.811447525	2.317549844	0.887884548
StdDevF0	1.078497895	2.74759046	0.751333614
Max dF0	3.085304688	1.211111339	1.472668297
Mean dF0	1.079211158	-1.597625791	1.659623152
dF0 Range	1.517467854	1.052387228	1.117292524
StdDevdF0	2.585356493	1.859113206	1.143930008
Max Int	1.03462185	1.044617315	1.019276762
Mean Int	1.02794861	1.045610033	1.033178061
Int Range	1.029211251	1.142897043	0.949404417
StdDev Int	0.989561789	1.419458103	0.992874148
Max dInt	1.219646585	0.703387521	1.067154224
Mean dInt	0.196349952	1.629129185	0.732939084
dInt Range	1.027048338	1.136606218	0.779266319
StdDev dInt	1.310939021	1.153795271	1.047416081
Jitter (local)	0.902560455	1.097261735	0.962660028
Shimmer (local)	0.843981481	0.906095679	0.946141975
Fraction of locally unvoiced frames	0.289218615	0.885618775	0.756851312
Mean harmonics-to-noise ratio (dB)	1.505347047	1.113684354	0.989987058
Speech Rate	1.455052002	1.298921153	1.160345416
End Frequency (Hz)	1.415737943	1.091533039	0.741638676

Table A.9: Average multiplication factor for acoustic attributes between neutral and emotional for character Frederik

Emotion	**Anger**
Max F0	1.398987257
Mean F0	1.210709341
F0 Range	1.547731536
StdDevF0	1.241029656
Max dF0	1.865671392
Mean dF0	-0.014529202
dF0 Range	1.313470011
StdDevdF0	2.220270054
Max Int	1.038961144
Mean Int	1.069813547
Int Range	1.290352744
StdDev Int	1.030493169
Max dInt	0.889904493
Mean dInt	1.462883259
dInt Range	0.900107957
StdDev dInt	1.064881864
Jitter (local)	1.090938883
Shimmer (local)	1.085355771
Fraction of locally unvoiced frames	0.68982659
Mean harmonics-to-noise ratio (dB)	1.381825273
Speech Rate	1.039568905
EndFrequency (Hz)	2.687566592

Table A.10: Average multiplication factor for acoustic attributes between neutral and emotional for character Tante

Happiness	**Recognized%**	**F0 Mean**	**F0 Range**	**Tempo**
Burkhardt and Sendlmeier [4]	62% (1/9)	0%	100%	+20% or ”-20%”
Cahn [5]	48% (1/6)	-3 reference line (=pitch value between accents) ”-8”, but less final lowering ”-4”	+10 contour slope ”+5” (I.e. pitch range is expanding throughout the utterance) accent shape (=steepness of F0 contour at accents) ”+10”	+2 fluent pauses ”-5”, hesitation pauses ”-8”
Heuft et al. [21]				very fast
Iriondo et al. [22]	untested	increased (10-50%)	increased (120%)	decrease of silence duration (20%)
Montero et al. [28, 29]	19% (1/5)	higher than neutral		pause duration: ca. half of neutral (both sentence-final and intra-sentence)
Murray and Arnott [34]	not recognized		+9 s.t.; reduce the amount of pitch fall at end of utterance by 10% of F0	+30 wpm; duration of stressed vowels +50%; modify phoneme durations for regular stressing (=¿ time between two stressed phonemes = 550ms or multiple thereof)
Murray et al. [33]	raised, high-pitched		slightly increased	
Tao et al. [53]	-0.38	37.20%	73%	-2.40%

Table A.11: Prosody rules for a synthesized happy voice, part 1. From [46] appendix A

Happiness	**Loudness**	**Voice Quality**	**Other**
Burkhardt and Sendlmeier [4]		modal or tense; ”lip-spreading feature”; F1/F2+10%	Wave pitch contour model: main stressed syllables are raised (”+100%”), syllables in between are lowered (”-20%”)
Cahn [5]		breathiness ”-5”brilliance ”-2”	stress frequency (no. of accents) ”+5”; precision of articulation ”-3”
Heuft et al. [21]			
Iriondo et al. [22]	stable intensity		fast inflections of tone; F0 rise and fall times similar, no high plateaux, pitch-energy relation asynchronous (energy peaks 100 to 150ms earlier than pitch peaks, sounds like laughter)
Montero et al. [28, 29]			
Murray and Arnott [34]	+3dB		eliminate abrupt changes in pitch between phonemes
Murray et al. [33]			
Tao et al. [53]	3.30%		Recognized number is calculated as the difference between the mean level labeles for the synthesized emotion minus the mean level labeled for the original speech. The number for the strong variant of the emotion is shown here.

Table A.12: Prosody rules for a synthesized happy voice, part 2. From [46] appendix A

Joy	**Recognized%**	**F0 Mean**	**F0 Range**	**Tempo**
Burkhardt and Sendlmeier [4]		50%	100%	30%
Mozziconacci [30, 32]	62% (1/7)	end frequency 155Hz (male speech)	excursion size 10 s.t.	duration rel. to neutrality: 83%

Table A.13: Prosody rules for a synthesized joyous voice, part 1. From [46] appendix A

Joy	**Loudness**	**Voice Quality**	**Other**
Burkhardt and Sendlmeier [4]		modal or tense; ”lip-spreading feature”; F1/F2+10%	Wave pitch contour model: main stressed syllables are raised (”+100%”), syllables in between are lowered (”-20%”)
Mozziconacci [30, 32]			final intonation pattern 1andA or 5andA; avoid final patterns A, EA and 12.

Table A.14: Prosody rules for a synthesized joyous voice, part 2. From [46] appendix A

Sadness	**Recognized%**	**F0 Mean**	**F0 Range**	**Tempo**
Burkhardt and Sendlmeier [4]	Crying despair 69% (1/9), Quiet Sorrow 38% (1/9)	despair +100%sorrow -20%	-20% pitch variability (within syllables) "-20%"	despair -20%sorrow -40%
Cahn [5]	91% (1/6)	0 reference line "-1", less final lowering "-5"	-5 steeper accent shape "+6"	-10 more fluent pauses "+5", hesitation pauses "+10"
Heuft et al. [21]		low		fast
Iriondo et al. [22]		decreased (10-30%)	decreased (30-50%), less than 30Hz	duration of silences increased (50-100%)
Campbell and Marumoto [7]	52% (1/3)	lower (-10Hz)	reduced (0.6875)	8% slower (duration * 1.08)
Montero et al. [28, 29]	67% (1/5)	lower than neutral		pause duration: ca. a third more than for neutral (both sentence-final and intra-sentence)
Mozziconacci [30, 32]	47% (1/7)	end frequency 102Hz	excursion size 7 s.t.	duration rel. to neutrality 129%
Murray and Arnott [34]	well recognized with neutral text	-30 Hz	-2. S.t.	-40 wpm
Murray et al. [33]		lowered		slightly slower
Rank and Pirker [41]	69% (1/4)	-3	0.5	duration of vowels: 1.4, voiced cons: 1.5, unvoiced cons. 1.2; pause duration 3.0, pause duration variability 0.5
Tao et al. [53]	-0.59	-13.90%	-30.90%	4%

Table A.15: Prosody rules for a synthesized sad voice, part 1. From [46] appendix A

Sadness	**Loudness**	**Voice Quality**	**Other**
Burkhardt and Sendlmeier [4]		breathy	F0 flutter (jitter): despair"FL 200"sorrow"FL 300"
Cahn [5]	-5	breathiness "+10", brilliance"-9"	stress frequency "+1" precision of articulation "-5"
Heuft et al. [21]			
Iriondo et al. [22]	decreased (10-25%)		null inflections of intonation; discourse fragmentation increased (10%)
Campbell and Marumoto [7]	intensity range reduced by 5%		
Montero et al. [28, 29]			
Mozziconacci [30, 32]			Final intonation pattern 3C; avoid final pattern 5andA
Murray and Arnott [34]	-2 dB	spectral tilt +65%	decrease articulation precision (reduce "a" and "I" vowels); eliminate abrupt changes in pitch between phonemes; replace upward inflections with downward inflections; add 80ms pause after each word with more than 3 phonemes.
Murray et al. [33]		a bit of artificial "laryngealisation"	
Rank and Pirker [41]	amp. Vowels 0.7, voiced cons. 0.7, unvoiced cons 0.8; amp. Shimmer 0.0	creaky rate 0.02, glottal noise 0.4	F0 jitter 0.0005; articulation precision 0.95
Tao et al. [53]	-6.80%		Recognized number is calculated as the difference between the mean level labeles for the synthesized emotion minus the mean level labeled for the original speech. The number for the strong variant of the emotion is shown here.

Table A.16: Prosody rules for a synthesized sad voice, part 2. From [46] appendix A

Anger	**Recognized%**	**F0 Mean**	**F0 Range**	**Tempo**
Burkhardt and Sendlmeier [4]	Hot Anger 29% (1/9); cold anger 60% (1/9)	hot +50%cold 20%		30%
Cahn [5]	44% (1/6)	-5 reference line ”-3”, extreme final lowering ”+10”	+10 steep accent shape ”+10”	+8 less fluent pauses ”-5”, hesitation pauses ”+-7”
Heuft et al. [21]		very low	narrow	
Iriondo et al. [22]	Fury		very wide (can exceed 140Hz)	slower; reduction of the number of silencces (25%); increase the duration of silences (10%)
Campbell and Marumoto [7]	65% (1/3)	raised (+7Hz)	wider (*1.125)	2% faster (duration * 0.98)
Montero et al. [28, 29]	Cold anger 7% (1/5)	like neutral	nearly no declination (final peaks as high as initial peaks)	pause duration: roughly 2/3 of neutral (both sentence-final and intra-sentence)
Mozziconacci [30, 32]	51% (1/7)	end frequency 110Hz	excursion size 10 s.t.	duration rel to neutrality 79%
Murray and Arnott [34]	well recognized with neutral text	+10Hz	+9 s.t.	+30 wpm
Murray et al. [33]	Anger	raised	increased	faster
Rank and Pirker [41]	40% (1/4)	0	2	duration of vowels 0.75, voiced cons 0.85, unvoiced cons 0.9, pause duration 0.8, pause duration variability 0.0
Tao et al. [53]	-1.14	23.30%	93.00%	-9%

Table A.17: Prosody rules for a synthesized angry voice, part 1. From [46] appendix A

Anger	**Loudness**	**Voice Quality**	**Other**
Burkhardt and Sendlmeier [4]		tense	cold: vowel articulation precision (formant overshoot) +30% for stressed syllables, -20% for unstressed syllables
Cahn [5]	10	breathiness ”-5”, brilliance ”+10”	precision of articulation ”+5”
Heuft et al. [21]			
Iriondo et al. [22]	raising intensity from the begin to the end (5-10dB)	most characteristic for fury: increase of energy in 500-636Hz and 2000-2500Hz bandwidths (10-15dB)	variation of the intonation structure (20-80Hz); pitch rises faster than falls; monosyllabic high pitch plateaux; downward declination (approx: topline 245-150Hz, baseline 190-90Hz)
Campbell and Marumoto [7]	intensity range increased by 10%		
Montero et al. [28, 29]			
Mozziconacci [30, 32]			Final intonation pattern 5andA or A or EA; avoid final pattern 1andA or 3C
Murray and Arnott [34]	+6 dB	laryngealisation +78%F4 frequency 175Hz	increase pitch of stressed vowels (2ary: +10% of pitch range, 1ary: +20%; emphatic: +40%)
Murray et al. [33]		artificial ”laryngealisation”	
Rank and Pirker [41]	amp vowels 1.3, voiced cons 1.2, unvoiced cons 1.1, amp shimmer 0.0		articulation precision 1.05
Tao et al. [53]	2.00%		Recognized number is calculated as the difference between the mean level labeles for the synthesized emotion minus the mean level labeled for the original speech. The number for the strong variant of the emotion is shown here.

Table A.18: Prosody rules for a synthesized angry voice, part 2. From [46] appendix A

Fear	**Recognized%**	**F0 Mean**	**F0 Range**	**Tempo**
Burkhardt and Sendlmeier [4]	Fear 52% (1/9)	150%	20%	30%
Cahn [5]	Scared 52% (1/6)	+10 reference line "+10", no (or negative?) final lowering "-10"	+10 steeply rising contour slope "+10" steep accent shape "+10"	+10 no fluent pauses "-10", but hesitation pauses "+10"
Heuft et al. [21]	Fear		narrow	very fast
Iriondo et al. [22]	Fear	increased (5-10%)	decreased (5%)	faster (decrease of duration of phonic groups by 20-25%); decrease duration of silences (10%)
Mozziconacci [30, 32]	Fear 41% (1/7)	end frequency 200Hz	excursion size 8 s.t.	duration rel to neutrality 89%
Murray and Arnott [34]	Fear (not recognized with neutral text)	+20Hz	+3 s.t.; end F0 baseline fall +100Hz	+20 wpm
Murray et al. [33]	Fear	even more raised than happy ("squeaky")		very much increased
Rank and Pirker [41]	Fear 18% (1/4)	1.5	2	duration of vowels 0.65, voiced cons 0.55, unvoiced cons 0.55, pause duration 0.6, pause duration variability 0.5
Tao et al. [53]	-0.92	-12.40%	-31.50%	-3.10%

Table A.19: Prosody rules for a synthesized fearful voice, part 1. From [46] appendix A

Fear	**Loudness**	**Voice Quality**	**Other**
Burkhardt and Sendlmeier [4]		falsetto	
Cahn [5]	10	brilliance ”+10” laryngealisation ”-10”	stress frequency ”+10”, loudness ”+10”
Heuft et al. [21]			
Iriondo et al. [22]	raised intensity (10%) energy globally rising		fast variations of pitch (by 60-100 Hz in 20-30ms); bi- or trisyllabic high pitch plateaux; high plateaux rising (approx: topline 180-250Hz, baseline 140-140Hz)
Mozziconacci [30, 32]			final intonation pattern 12; avoid final pattern A or EA
Murray and Arnott [34]		laryngealisation +50%	increase articulation precision (unreduce vowels; duration of plosives +30%)
Murray et al. [33]			
Rank and Pirker [41]	amp vowels 1.2, voiced cons 1.0 unvoiced cons 1.1 amp shimmer 0.05	creaky rate 0.003; glottal noise 0.5	F0 jitter 0.35; articulation precision 0.97
Tao et al. [53]	-5.10%		Recognized number is calculated as the difference between the mean level labeles for the synthesized emotion minus the mean level labeled for the original speech. The number for the strong variant of the emotion is shown here.

Table A.20: Prosody rules for a synthesized fearful voice, part 2. From [46] appendix A

Surprise	**Recognized%**	**F0 Mean**	**F0 Range**	**Tempo**
Cahn [5]	Surprised 44% (1/6)	0 reference line ”-8”	+8 steeply rising contour slope ”+10” steeper accent shape ”+5”	+4 less fluent pauses ”-5” hesitation pauses ”-10”
Iriondo et al. [22]	Surprise	increased (10-15%)	increased (15-35%) high inflections in intonation	faster (decreased duration of phonic groups by 10%)
Montero et al. [28, 29]	Surprise 76% (1/5)	much higher than neutral (ca 200Hz for male speaker)		pause duration: ca 60% of neutral (both sentence-final and intra-sentence)

Table A.21: Prosody rules for a synthesized surprised voice, part 1. From [46] appendix A

Surprise	**Loudness**	**Voice Quality**	**Other**
Cahn [5]	5	brilliance ”-3”	
Iriondo et al. [22]	increased (3-5 dB)		
Montero et al. [28, 29]			

Table A.22: Prosody rules for a synthesized surprised voice, part 2. From [46] appendix A

Neutral	**Recognized%**	**F0 Mean**	**F0 Range**	**Tempo**	**Other**
Burkhardt & Sendlmeier [4]	Neutral 55% (1/9)				
Heuft et al. [21]	Neutral	low	narrow	slow	
Montero et al. [28, 29]	Neutral 76% (1/5)				
Mozziconacci [30, 32]	Neutral 83% (1/7)	end frequency 65Hz	excursion size 5 s.t.		final intonation pattern 1&A; avoid final patterns 3C and 12

Table A.23: Prosody rules for a synthesized neutral voice. From [46] appendix A

Boredom	**Recognized%**	**F0 Mean**	**F0 Range**	**Tempo**
Burkhardt and Sendlmeier [4]	Boredom 71% (1/9)	-20%	-50% reduced pitch variability (within syllables) ”20%” (=> almost flat intonation contour)	-20% additional lengthening of stressed syllables (40%)
Mozziconacci [30, 32]	Boredom 94% (1/7)	end frequency 65Hz	excursion size 4 s.t.	duration rel to neutrality 150%

Table A.24: Prosody rules for a synthesized bored voice, part 1. From [46] appendix A

Boredom	**Loudness**	**Voice Quality**	**Other**
Burkhardt and Sendlmeier [4]		modal	reduced vowel precision (formant undershoot) stressed syllables ”-20%”, unstressed syllables ”-50%”
Mozziconacci [30, 32]			final intonation pattern 3C; avoid final patterns 5andA and 12

Table A.25: Prosody rules for a synthesized bored voice, part 2. From [46] appendix A

A.3 Evaluation Experiment

Alleen trager en minder luid, ipv emotie
de droefenis komt denk ik meer tot uitdrukking in tempowisselingen: ruimte na waar bv (heeft ook met uitstraling van vermoeidheid te maken)
Het klinkt als een langzamer afgespeelde versie van de neutrale. In het woord ”vandaan” zit te veel toonhoogtevariatie, die maakt dat het raar klinkt. Doordat het zo langzaam wordt afgespeeld klinkt het meer als een achterdochtige vrouw dan als een bedroefde.
het klonk eerder moe dan bedroefd
klinkt in elk geval meer bedroefd dan ”standaard”... Maar klinkt wel heel erg monotoon.
vooral omdat de zin rustiger wordt uitgesproken klinkt het bedroefd
ze zegt VERhaald volgens mij...

Table A.26: Comments on the sad sentence 1: ”Waar heb je dat vandaan gehaald?”

het is precies hetzelfde als de vorige, alleen dan een andere tekst... ’k mis het echte bedroefde erin.. maarjah...
het klinkt vooral suf
Het woordt ’moeten’ begint te hoog en valt daardoor uit de toon met de rest van de zin.
hooguit zou de klemtoon meer op ’dat’ kunnen liggen
Ik vind het accent op ”niet” nogal vreemd (ik zou accent op ”doen” verwachten) en daardoor klonken ze allebei wat raar. NB: eigenlijk vind ik dit fragment wel bedroefd(er) klinken, net als het vorige
toonhoogte van de laatste 2 woorden mag wel verschillend zijn, doen lager dan moeten.
Wat betere variaties in woordlengtes

Table A.27: Comments on the sad sentence 2: ”Dat had je niet moeten doen.”

aan het einde van de zin moet het omhoog gaan
als je verrast ben dan ga je volgens mij op het eind van de zin omhoog, en in het fragment eindigt de zin laag.
Het heeft wel iets weg van verrast, maar het kan ook geagiteerd zijn doordat er nog veel nadruk op het woord ’niet’ ligt.
jammer dat je er geen schrikademhaling bij kunt doen, voor de zin, een inademingsgeluid bv.
Klinkt eerder boos dan verrast. Opnieuw: gekke plaatsing van het accent. (Deze plaatsing leidt bovendien eerder tot een boze interpretatie!)
spreeksnelheid zou iets rustiger kunnen
variatie in spreeksnelheid tijdens de zin zou misschien nog wat toe kunnen voegen

Table A.28: Comments on the surprised sentence 3: ”Dat had je niet moeten doen.”

het boze zit ’m in de accenten, in dit geval de variatie in geluidssterkte. ik zou meer nadruk leggen op ’best zelf’. daar zou je ook iets kunnen vertragen.
het klinkt eerder sjachereinig/boos
Het tweede deel ’kunnen verzinnen’ staat teveel los van het eerste gedeelte. in het eerste komt meer de dominantie van boos naar voren, maar in het tweede gedeelte ontbreekt dat.
Ik vind het vooral verontwaardigd klinken (maar dat is immers ook een vorm van boos)
klinkt eerder bedroefd
Klinkt geirriteerd en uit de hoogte, iets wat bij boosheid wel kan kloppen ja.
Klinkt verontwaardigd

Table A.29: Comments on the angry sentence 4: ”Dat had je best zelf kunnen verzinnen.”

Het deel ’schoon’ uit ’schoonfamilie’ is te hoog en maakt dat de hele zin raar klinkt. Het klinkt ook meer als een statement wat aan een dombo uitgelegd wordt dan als en bedroefd persoon.
het klinkt meer mistroostig dan bedroefd. Het gaat wel de bedroefde kant op. Laatste gedeelte van de zin is qua toonhoogte te vlak.
Het klinkt wat teneergeslagen, alleen ’schoonfamilie’ zit niet lekker in elkaar.
Ik vind het eerder sarcastisch klinken dan bedroefd. Maar dat ligt volgens mij niet aan bovenstaande aspecten. Eerder door de manier waarop ”mijn schoonfamilie” wordt uitgesproken; beetje overdreven gearticuleerd.

Table A.30: Comments on the sad sentence 5: ”Morgen komt mijn schoonfamilie op bezoek.”

Een beetje paniekerig, dus wel bang ja.
Het tweede gedeelte van de zin gaat qua toonhoogte naar beneden zoals het een standaard zin betaamd. Dan komt het normaler over. Het eerste gedeelte vind ik wel goed.
het was iets te langzaam
klemtoon op dat en niet, doen laten zekken in toonhoogte.

Table A.31: Comments on the scared sentence 6: ”Dat had je niet moeten doen.”

de nadruk ligt verkeerd. Als je deze zin liefdevol uitspreekt ligt de nadruk zeker op ’schat’
Fragment 2 klinkt net als fragment 1. Er moet meer nadruk op het woord ’schat’ en dat moet met meer nadruk en een zuchtje uitgesproken worden. (jaja, alsof zo’n automatisch ding dat kan..) Hij is ook te snel, het kan niet liefdevol zijn omdat er geen gedachte (=tijd) achter zit.
Heb je er uberhaupt wel wat aan verandert? Het klinkt bijna hetzelfde... alleen bij het woordje ’een’ zit een rare vervorming. Zoiezo zou ik de nadruk niet op ’toch’ leggen zeker niet bij een emotie als liefdevol.
Meer specifiek voor deze zin zou ik zeggen: sterker accent op ”ben”, en ”schat” langer aanhouden.
snelheidsvariatie in de zin maakt emotie: in deze zin kan ’ben je toch’ sneller uitgesproken worden en iets hoger dan de rest.

Table A.32: Comments on the loving sentence 7: ”Wat ben je toch een schat.”

agitatie komt wel naar voren, misschien moet je dat verrast noemen?
Beetje hyper, opgewonden (misschien eerder dan verrast)
ik denk dat de klemtonen verkeerd liggen, het ligt nu op zowel 'morgen' als 'schoonfamilie', terwijl, als je echt verrast bent, de klemtoon alleen op 'schoonfamilie' legt. en de woorden worden te snel uitgesproken.
Ja, maar spreeksnelheid misschien net iets te hoog
Klinkt eerder boos, omdat het snel uitgesproken wordt en de woorden zijn erg afgebeten.
maar de klemtoon moet wat sterker
meer opgewonden
morgen komt mijn schoonfamilie op bezoek mi mi mi mi sol fa mi re do
naar het einde toe van de zin kan het iets rustiger
Wederom gaat de intonantie van de zin in het tweede gedeelte naar beneden wat het verrassings element eruit haalt.

Table A.33: Comments on the surprised sentence 8: "Morgen komt mijn schoonfamilie op bezoek."

de stem valt hier wat weg, das erg jammer
de toonvariatie in de laatste 4 worden is te klein. kwijt kan meer benadrukt worden door iets meer tijd voor dat woord te nemen.
Het klinkt wel wat banger, maar misschien een optie om ook de nadruk op kwijt te leggen? Daar gaat het namelijk op en als je bang bent leg je daar onbewust de nadruk op... (denk ik)
nadruk ligt wat veel op 'niet'. Is bij andere fragmenten ook zo
ook hier weer de klemtonen opmerking, ze zegt: ik hoop NIET dat we hem kwijt zijn geraakt, maar ik denk dat je normaalgesproken zou zeggen: ik hoop niet dat we hem KWIJT zijn geraakt. En ze slikt het einde van de zin in, maar verder klinkt het wel wat paniekerig."
Paniek! Bezorgdheid.
zelfde als vorige, klemtoon sterker

Table A.34: Comments on the scared sentence 9: "Ik hoop niet dat we hem kwijt zijn geraakt."

Dat had je best zelf kunnen verzinnen. klemtoon, afknijping op einde van het laatste woord.
het zal alleen bedroefd klinken als de stem verder 'normaal' is.
iets te traag
Ja, maar spreeksnelheid net iets te traag.
klink moe (alweer)
Klinkt als iemand die levensmoe is, maar niet bedroefd. Bij bedroefdheid moeten woorden minder helder uigesproken worden. Er ligt ook veel nadruk op het woord 'zelf'.

Table A.35: Comments on the sad sentence 10: "Dat had je best zelf kunnen verzinnen."

blijheid toon is goed, echter ontbreekt de klemtoon op de ”vandaan” om aan te geven dat het een vraag is.
door de hoogte wordt het een totaal andere stem. De bedoeling is m.i. om de 2e stem blij te laten klinken (waarschijnlijk iets te hoog)
eerder bang
hehehe ’t klinkt net als een kabouter die iets vraagt :)
klemtoon op dat (hoger), vraagteken krijgt kleur door ’haald’ hoger te maken.
Klinkt eerder paniekerig en haastig dan blij. Juist bij zo’n ’domme’ zin als dit verwacht je een langerektere ’vandaan’ of ’waar’. De ’ge’ van ’gehaald’ is afleidend evenals het omhooggaan van de ’daan’ in ’vandaan’.
klinkt eerder verrast dan blij
ze zegt echt VERhaald, niet GEhaald ;) ik denk dat als je blij bent, dat je dan juist harder gaat praten, niet zachter.

Table A.36: Comments on the happy sentence 11: ”Waar heb je dat vandaan gehaald?”

bang word in het totaal niet goed uitgebeeld
de standaard arzelingen van angst komen niet over
deze klinkt meer verdrietig eigenlijk...
Klinkt eerder bezorgd dan echt bang. Sowieso is ook in het neutrale fragment de nadruk op ’vandaan’ te groot. Het neutrale fragment klinkt zelfs al enigzins beschuldigend. Dat verdwijnt niet geheel door het versnellen.
Bij echte angst krijgen mensen vaak ook een soort spraakgebrek, zoals niet uit de woorden kunnen komen of stotteren.”

Table A.37: Comments on the scared sentence 12: ”Waar heb je dat vandaan gehaald?”

De tweede klinkt een beetje raar, ik weet miet waar het aan ligt
er komt geen verandering in erg jammer
Het einde gaat wederom te snel en te veel naar beneden... Verder netjes...
in de 2e zin is de klemtoon op schoonfamilie niet prettig voor het gehoor. Ik ervaar niet iets liefdevols in deze zin
In het woord ’schoonfamilie’ zit een toonhoogte-hobbel waardoor het woord idioot klinkt. Verder denk ik bij liefdevol meer aan zacht gefluister en wat minder strak uitgesproken woorden.
Klinkt beetje raar.
misschien heeft het een liefdevol effect als je behalve schoonfamilie ook de woorden op bezoek hoger doet. ’zoek’ dan wel wer iets laten zakken.
ze klinkt schor in het woord ’schoonfamilie’, maar liefdevol kan ik dat niet noemen.

Table A.38: Comments on the loving sentence 13: ”Morgen komt mijn schoonfamilie op bezoek.”

Het woord ’eind’ moet langer gerekt worden om goed verveeld te klinken, evenals het woord ’wanneer’. Verder doet de snelheid (traagheid eigenlijk) wel een hoop.
ik denk dat verveling in snelheidsvariatie het beste uitgedrukt kan worden. het woord wanneer kun je misschien wat uitrekken, de woorden ’eind aan’ op dezelfde hoogte (misschien had je dat al)
Kan je ’wanneer’ ook wat meer uitrekken en ’aan’ omhoog laten aflopen?
klemtoon, als je verveeld bent dan zeg je: wanneer komt HIER nou eens een eind aan. en dan het liefst tegelijk met een zucht. het fragment klinkt enorm elektronisch.
Lijkt veel op bedroefd
maar de klemtoon op eind moet beter
Zou ook bedroefd kunnen zijn.

Table A.39: Comments on the bored sentence 14: ”Wanneer komt hier nou eens een eind aan.”

dat een beetje uitrekken in tijd, doen lager dan het woord ervoor.
Volgens mij is er alleen verandering bij het woordje 'had'...
Wederom weinig verschil tussen de 2 fragmenten. Liefdevol moet vooral zachter, minder scherpe woorden.

Table A.40: Comments on the loving sentence 15: "Dat had je niet moeten doen."

Er moet meer nadruk op het woord 'dat'. 'Dat' en 'wacht' van 'verwacht' moet meer omhoog.
het klinkt wel verrast, maar het eind van de zin wordt ingeslikt, en ik zou de klemtoon op het woordje DAT leggen, niet op het woordje NIET.
hij had iets sneller gemogen
Ik denk dat de klemtoon ook grotendeels de emotie bepaald. Bij 'verrast' verwacht je een klemtoonverplaatsing. Bij deze zin zou ik de klemtoon verplaatsen naar 'dat'.
klemtoon verkeerd
Misschien idee om de klemtoon op 'dat' te leggen?
moet aan het einde omhoog (toonhoogte)
nadruk ligt weer op het verkeerde woord.
toonhoogte van 'dat' mag hoger, 'sacht' lager. neem iets meer tijd voor 'dat'
Zou wel iets langzamer mogen. Opnieuw een gek accent. Ik zou het accent op "dat" leggen. "Maar DAT had ik niet verwacht!"

Table A.41: Comments on the surprised sentence 16: "Maar dat had ik niet verwacht."

Eerder boos, verontwaardigd.
Het woord 'schoonfamilie' heeft te veel toonhoogtevariatie. De stem zou ook iets meer 'afgeknepen' mogen klinken.
Klinkt een beetje raar op het einde maar verder wel goed.
laat de toonhoogte aan het eind nog iets zakken.
misschien stemgeluid iets aan de hoge kant

Table A.42: Comments on the scared sentence 17: "Morgen komt mijn schoonfamilie op bezoek."

andere zin was misschien beter geweest, iets van: "had je dat echt moeten doen?" met klemtoon op echt
De 'doen' moet naar beneden in plaats van omhoog. Verder mag er een zucht door de hele zin klinken en voor het woord 'niet' mag meer tijd genomen worden.
doen lager maken. verveeldheid komt pas naar voren bij toevoeging van een woord als 'nou' bij 'nou niet moeten'.
het is wel slaapverwekkend, maar het klinkt niet verveeld, eerder als iemand die net uit bed komt, of nog onder invloed is ofzo..

Table A.43: Comments on the bored sentence 18: "Dat had je niet moeten doen."

die ’verrast’ emotie is wel ok, behalve dat de woorden ingeslikt worden en dat de klemtonen verkeerd liggen.
Eerder boos; bij verrassing zou het accent op ”dat” liggen. (Of desnoods op ”vandaan”.) Met de andere aspecten van de intonatie is niets mis
in de context van een verhaal kan dit verrast genoeg zijn. toch mis ik de variatie in tijd:waar meer tijd geven, dat en ’haald’ iets hoger.
klemtoon op dat moet sterker
Klinkt als een moeder die haast heeft en NU wil weten waar haar kind dat snoepje vandaan heeft gehaald. Geagiteerd en haastig dus, niet verrast. De ’dat’ moet met meer volume, hoger en langerekter worden uitgesproken.
Misschien net iets te snel.

Table A.44: Comments on the surprised sentence 19: ”Waar heb je dat vandaan gehaald?”

dit is te hoog en klinkt voor mij meer als ’bang’
Eerder angstig. Heel ”geknepen”.
eerder bang
het lijkt wel paulus de boskabouter! als je blij bent krijg je niet opeens een heel hoog beknepen stemmetje volgens mij.
Kan blij en bang zijn.
Klinkt als iemand met een dichtgeknepen strot of na een zakhak != blij. De nadruk op ’niet moeten’ geeft het nog wel wat van een beschuldiging. Het woordje ’doen’ mag wat langer. Er moet ook een ondertoon van niet-menen onder zitten, maar dat is lastig en context-gericht: je bent waarschijnlijk blij omdat die persoon het wel heeft gedaan. ”niet” kan hoger, ”doen” lager in toonhoogte.

Table A.45: Comments on the happy sentence 20: ”Dat had je niet moeten doen.”

De ’daan’ van ’vandaan’ moet langer en niet zo hoog. De laatste lettergreep van ’gehaald’ mag ook langer. Meer nadruk op ’dat’ zodat het meer klinkt als ’dat nou weer’. Verder is de snelheid wel ok en de hoogte ook.
Klinkt wel verveeld maar zou iets minder langzaam mogen
neem meer tijd voor ’waar’ en ’dat’, ’dat’ mag nog iets hoger, ’haald’ weer wat laten zakken.
Ook hier zou ik bij ’verveeld’ een klemtoonverschuiving of eigenlijk helemaal geen klemtoon. Verveelde mensen spreken heel monotoon.

Table A.46: Comments on the bored sentence 21: ”Waar heb je dat vandaan gehaald?”

doen weer wat lager in toonhoogte.
Het eerste gedeelte vind ik goed klinken. ”dat had je” daarna lijkt het alsof iemand anders het aan het inspreken is...
het klinkt meer gefrustreerd, als dat je nabij een zenuwinzinking bent, zo’n overspannen leraar voor de klas, je kent het wel ;)
Hier zou ik verwachten er meerdere duidelijke klemtonen voorkomen. En meer staccato, pauzes tussen de woorden.
klemtoon op niet sterker
’niet’ mag nog hoger = meer nadruk. Verder is hij goed. Het onbewerkte fragment klinkt ook al ietwat geirriteerd.

Table A.47: Comments on the angry sentence 22: ”Dat had je niet moeten doen.”

te snel
als een gespannen elastiekje, mag lager en rustiger. laat de hoogte van ’zoek’ eens zakken en neem iets meer tijd voor ’morgen’
De laatste lettergreep van ’bezoek’ mag langer, in anticipatie. ’schoonfamilie’ mag meer nadruk. Ook meer nadruk op ’morgen’, ook in anticipatie. Verder is het fragment veel te hoog, net een muis. En van opwinding (= blijheid) mag men harder gaan praten.
kon het slecht verstaan
Zou van alles kunnen zijn (boos, gedecideerd) maar niet blij. Weet niet waar het aan ligt.

Table A.48: Comments on the happy sentence 23: ”Morgen komt mijn schoonfamilie op bezoek.”

alleen weinig intonantie...
behalve als het eerste fragment ook liefdevol is ;)
deze zijn m.i. weer hetzelfde. dus geen vergelijk mogelijk
Door de nadruk op de vraag-intonatie klinkt het als een beschuldiging. Liefdevol mag zachter, minder strak uitgesproken.
ik weet niet of je zo’n zin liefdevol kunt uitspreken. het is nu te monotoon. of je met klemtonen op ’waar’(sterkte) en ’dat’(hoger) wel deze emotie krijgt vraag ik me af.

Table A.49: Comments on the loving sentence 24: ”Waar heb je dat vandaan gehaald?”

Das wel heeeel mild...
’dat’ zou je nog iets hoger in toon en sterker kunnen doen, ’haald’ ook wat optrekken om het mider saai te laten klinken.
Klinkt wel tamelijk boos, komt ook doordat de vraag al als een beschuldiging klinkt. Het woord ’waar’ mag harder en langer voor meer nadruk. Verder is de snelheid wel goed. De 2e lettergreep van ’vandaan’ krijgt nog iets te veel nadruk doordat ie omhoog gaat.
ze klinkt eerder bang voor het mogelijke antwoord dat ze krijgt, dan dat ze boos is. ik denk dat als je boos bent dat je stem dan lager en harder wordt, om mee te beginnen.

Table A.50: Comments on the angry sentence 25: ”Waar heb je dat vandaan gehaald?”

te snel gesproken
beetje te hoge toonhoogte
klinkt ook al blij als het iets rustiger en lager is.
Maar misschien net iets te snel tempo.
Meer lengte in ’geloooooopen’, anticipatie. ’af’ van ’afgelopen’ mag vrij hoog beginnen. Verder klinkt het meer als die mier uit de Fabeltjeskrant dan als een blij persoon.
neem meer tijd voor ’bijna’, je kunt dat ook iets hoger laten klinken. dan zit er meer opluchting in!

Table A.51: Comments on the happy sentence 26: ”Deze test is bijna afgelopen.”

De woorden 'morgen' en 'schoonfamilie' mogen lager en harder dan de rest van de zin = dreigender. Verder is het woord 'schoonfamilie' gewoon weer raar met te veel hoogte-huppels.
hierbij hetzelfde wat ik ook al eerder genoemd heb nl. de klemtoon in schoonfamilie.
Ja, maar toonhoogte variatie is een beetje vaag.
maar toch neeeet niet
wel vreemd, die toonhoogte op 'schoon'. ik denk dat je meer bereikt door het iets luider te laten klinken. misschien kun je 'morgen' iets sterker (luid) maken.
zo! die is echt gefrustreerd! wat voor schoonfamilie zou dat zijn?
Zou boos kunnen zijn, maar sommige van de eerdere fragmenten klonken bozer dan deze... Er zitten ook wat rare artefacten in

Table A.52: Comments on the angry sentence 27: "Morgen komt mijn schoonfamilie op bezoek."

een beetje een "wat maakt dat nou weer uit" toon :)"
Heel mooi die langerekte 'schoon'. Ondanks de blikkerigheid van de stem klinkt dit wel verveeld.
iets te traag. Op zich klinkt het aardig verveeld
Ja, maar Spreeksnelheid misschien net iets te langzaam.
snelheid is wel ongeveer goed.je moet toch meer varieren in de snelheid over de hele zin genomen.
te langzaam
wel erg overdreven (met name de snelheid)

Table A.53: Comments on the bored sentence 28: "Morgen komt mijn schoonfamilie op bezoek."

Appendix B

Maintainer Manual

B.1 Festival code manual

Festival uses a very adaptable and modular approach in its text-to-speech conversion. The source-code is partly in c++ and partly in scheme (specifically SIOD), a lisp variant. This makes tracing the code path more difficult as calls are made from scheme code to c++, from c++ to scheme, and internally called lisp instructions in c++. Furthermore, Festival uses libraries from the Edinburgh speech tools which contains the extended SIOD implementation. SIOD stands for Scheme in one defun. The original SIOD implementation has been extended by Alan Black to include features used in the Edinburgh speech tools (EST) and Festival. Documentation on festival is available at http://festvox.org/docs/manual-1.4.3/ Documentation on Scheme is available at http://tinuviel.cs.wcu.edu/res/ldp/r4rs-html/ The full list of function- and variable descriptions extracted from the source code is available on the digital attachment.

B.1.1 Source code locations

It is assumed that Festival - and the Edinburgh Speech Tools directories are located in the same directory. For instance Festival could be located in /usr/src/festival and the EST in /usr/src/speech_tools. Of course symlinks are ok as well. The SIOD implementation is located in speech_tools/siod. The Festival c++-based modules are located in festival/src/modules. The Festival scheme files are located in festival/lib. The NeXTeNS source-code additions are located in festival/lib/net_nl, festival/voices/dutch, festival/src/modules/NextensTimbl and festival/src/modules/NextensMbt.

B.1.2 Compiling instructions

Compile EST first, then compile festival. Note that the NeXTeNS modules aren't properly cleaned on a 'make clean'.

B.1.3 Code structure

Each part, from the input parser, the tokenizer, syntax checker, POS-tagger to the duration and intonation modules is listed in the utterance type specifications. These specifications can be overridden by voice initialisation routines, like net_nl_ib_mbrola (nextens' default voice).

It all starts with a call to tts_file, which starts an input parser that scans the input, puts it into an utterance object, which is then passed on to actual synthesising function (starting in tts.scm, then text.cc and text_modes.cc and ending again in tts.scm with the calling of tts_hooks). The tts_hooks then calls utt.synth (synthesis.scm) which in turn calls all the functions in the utterance type list.

The nextens "Text utterance type" consists of a call to the following functions (set in net_nl_synthesys):

Initialize	(in festival/src/modules/base/modules.cc)
Text	(in festival/src/modules/Text/text.cc)
Token	(in festival/lib/token.scm)
POS	(in festival/lib/pos.scm)
Syntax	(in festival/lib/net_nl/net_nl_syntax.scm)
Phrasify	(in festival/lib/phrase.scm)
Intonation	(in festival/lib/intonation.scm)
Tune	(in festival/lib/net_nl/net_nl_tune.scm)
Word	(in festival/lib/lexicons.scm)
Pauses	(in festival/lib/pauses.scm)
PostLex	(in festival/lib/postlex.scm)
Duration	(in festival/lib/duration.scm)
Int_Targets	(in festival/lib/intonation.scm)
Wave_Synth	(in festival/lib/synthesis.scm)

Table B.1: NeXTeNS Text utterance type

The festival functions in the list above (being those not in the net_nl directory) are configured via an extra parameter, which contains the method to use in each module. These parameters are set in festival/lib/voices/dutch/net_nl_ib_mbrola/festvox/net_nl_ib_mbrola.scm. The parameters refer to functions which are then called from the functions in the above list.

The list of parameters from the net_nl_ib_mbrola voice is as follows:

Token_Method	Token_Any	(in festival/src/modules/Text/token.cc)
Syntax_Method	nil	
Phrasify_Method	nl::prosit-break	(in festival/lib/net_nl/net_nl_break_prosit.scm)
Int_Method	nl::prosit_accent_placement	(in festival/lib/net_nl/net_nl_accent_prosit.scm)
Tune_Method	nl::basic_tune_choice	(in festival/lib/net_nl/net_nl_tune.scm)
Word_Method	nl::word	(in festival/lib/net_nl/net_nl_lex.scm)
Pause_Method	nl::pauses	(in festival/lib/net_nl/net_nl_pauses.scm)
Duration_Method	KUN_Duration	(in festival/lib/net_nl/net_nl_dur_kun.scm)
Int_Target_Method	ToDI-intonation	(in festival/lib/net_nl/net_nl_int_todi.scm)
Synth_Method	nl::mbrola_synth	(in festival/lib/net_nl/net_nl_mbrola.scm)

Table B.2: List of parameters from net_nl_ib_mbrola

B.1.4 Data structure

Throughout the TTS process, a single data object - the utterance object - is passed along to almost all the functions. The functions read from, modify and add to the data of this utterance object. This object contains multiple trees with EST_Value objects. An utterance can contain many "Relations", which are sets of trees. Each unique EST_Value object can only occur in a relation once, but it can occur in multiple relations simultaneously. An item can have zero or more key-data pairs called "features".

The Initialize function creates an utterance object with just an empty Token relation, which is then filled with tokens read from the parsed text in the Text function. The tokens are then tokenized and translated into words which are stored in the Word relation. In order to link the words to the token(s) they were derived from, they are also added as subnodes to the token in the Token relation. The nl::prosit-break function adds "pbreak" features to the items in the Word relation which are translated in KUN_Duration into pause segments for the MBROLA output.

The nl::word function creates the NeXTeNS-specific trees in the utterance object, being: Syl-Part, Syllable, Foot, ProsWord1, ProsWord2, Word-Pros, ProsTree and Syl-Int. Each word in the Word relation is also in the ProsWord1 relation. Compound words (samengestelde woorden) are split up and each is put in the ProsWord2 relation. Each ProsWord2 has Appendix or Foot subnodes, which in turn consist of Syllable subnodes. Syllable nodes have one or more from the set Onset, Nucleus and Coda as subnodes (pending on the type of syllable), which in turn each have one Segment subnode. The segment is directly linked to the MBROLA output that festival passes on to the MBROLA synthesiser. The whole tree, from ProsWord1 to Segment is stored in the ProsTree relation.

The ToDI-intonation function translates the ToDI accents present in the features of items in the Intonation relation into F0 targets of items in the Segment relation.

The nl::mbrola_synth function reads the data from the Segment relation and feeds it to the MBROLA engine. The resulting wave file is then reloaded into festival.

B.1.5 SSML implementation

The XMLreader checks the SSML_elements variable for every tag encountered. If it is a <prosody> or <break> tag, they will be stored in the SSMLTree relation, with the tokens to which the tag applies as subnodes. Because a tag can apply to multiple tokens, and a token can be influenced by multiple tags, a single tree is insufficient to store this (because a token can be in multiple trees in the same relation, which is not allowed). To this end, when a token is already in the SSMLTree, a copy is made and added instead. In order to keep track of which copy belongs to which original, all tokens added to the SSMLTree are also added to the LookupTree relation, with original tokens as root elements and their copies as subnodes. This allows an easy iteration of all recognised SSML tags and which tokens they apply to (and vice versa).

The various implemented SSML tags influence different prosodic attributes: speech-rate, F0, F0 range and volume. There are a few places in the above functions where code from ssml-mode.scm is called in order to apply the information present in the SSML tags to the utterance data. Calling functions like this is done via hooks. Calls to the function 'apply_hooks' are made, with a list of functions as argument, in specific places. This is the list of hooks for the SSML handling (name and function called)

post_phrasing_hooks	nl::ssml_apply_break_changes
nl::post_dur_kun_hooks	nl::ssml_apply_duration_changes
nl::during_ToDI-intonation_hooks	nl::ssml_apply_range_changes
nl::post_ToDI-intonation_hooks	nl::ssml_apply_intonation_changes
after_analysis_hooks	ssml_apply_volume_changes

Table B.3: List of SSML hooks

The names of the hooks speak for themselves: post_phrasing_hooks are called after the Phrasify function, nl::post_dur_kun_hooks are called after KUN_Duration, etc. After_analysis_hooks are called before the actual MBROLA synthesis, but after everything else. The reason that the F0 range changes are done during the ToDI-intonation process is because the F0 is calculated on the fly based on a global parameter N. This was needed in order to be able to change the F0 range for just the words encapsulated by the <prosody>< /prosody> tags and without interfering too much with the existing code.

B.2 VST XML Preparser

The Preparser converts XML from the VST System into SSML. This

B.3 SSML Function list

In this section, a list of the functions present in ssml-mode.scm is iterated, along with the arguments expected and a short description on what the function does/returns.

Table B.4: SSML Function list

Function name	Arguments	Comment
ssml_init_globals	none	Initialises global variables used in ssml processing.
css2_time_to_ms	text	Returns the amount of time specified in according to css2-specification in ms
ssml_init_func	none	Initialisation for SSML mode
ssml_exit_func	none	Exit function for SSML mode
ssml_setup_voice_params	none	Set up original values for various voice parameters.
ssml_select_language	ATTLIST	SSML mode language selection. Translates xml:lang parameters to languages.
print_item	item	Prints EST_Item.
print_list	list	Recursively prints out anything in list.
ssml_print_item_recursive	item level	Recursive version of print_item. level contains the prefix used for indenting child items.
ssml_print_utt_relation	UTT relation	Prints all items in given relation.
ssml_push_language	language	Adds language to the top of the stack. Sets the next voice to be activated to default.
ssml_pop_language	none	Removes the top language from the stack. It is up to the caller to also call ssml_pop_voice.
ssml_push_voice	voice	Adds voice to the top of the stack and activates it. The top of the stack is the car element.
ssml_pop_voice	none	Removes the top voice from the stack and activates the prior voice. If there is no prior voice, the default voice for the current language is activated.
ssml_mark_tokens	utt type	Marks all tokens from the last on backwards to the starttoken. The type of mark given has to match the last item on the SSMLStack.
ssml_add_SSMLStack_item	UTT item features	Adds an item with name itemname and features to the SSMLStack.
ssml_print_SSMLStack	UTT	Prints the SSMLStack associated with the utterance.
ssml_pop_SSMLStack_item	UTT itemname	Pops item with itemname from the top of the SSMLStack if it's there.
nl::ssml_get_word_segments	utt word	Returns a list of segment items (EST_Val items) belonging to the word (item of 'Word relation) of the segment.

Continued on next page

Table B.4 – continued from previous page

Function name	Arguments	Comment
nl::ssml_apply_break_changes	utt	Walks through the 'Word tree and adds break features to words when indicated by ssml tags.
nl::ssml_change_segment_durations	segments duration	Expects duration to contain a list ("absolute",value) or ("relative',value) and segments to contain a list of segments and applies the duration to all segments.
nl::ssml_sum_segment_durs	number segments	Returns the total sum of the 'dur feat of all segments in the given list.
nl::ssml_fix_segment_endtimes	utt	Walks through the segments of the 'Segment relation and recalculates endtimes from the segments' 'dur feat.
nl::ssml_apply_duration_changes	utt	Walks through the 'Word tree and gathers prosody tags. Then applies the following speechrate-related tags: FIXME and FIXME.
nl::ssml_interpolate_target_f0	utt target_segment	Interpolates the f0 for the given segment from the surrounding segments with a target pitch. Interpolation method used is linear over the segments time (not the amount of segments). Returns interpolated segment f0.
nl::ssml_change_segment_pitch	utt segments pitch	Updates the pitch for all segments. If a segment has no pitch but needs a change, it is obtained through linear interpolation from surrounding segments with pitch. pitch can have the following values: ("absolute", value) with value in Hz, ("relative", value) with value a multiplication factor, ("shift", value) with value in Hz.
ssml_get_previous_segment	utt segment	Returns the segment prior to segment, even if it is not in the same word.
ssml_get_next_segment	utt segment	Returns the segment after segment, even if it is not in the same word.
ssml_parse_pitch_attribute	attribute	Checks attribute W3C compliance and returns a modification tuple. Modification tuple is ("absolute",value) with value in Hz, or ("relative",value) with value a multiplication factor or ("shift",value) with value in Hz.
ssml_parse_contour_attribute	attribute	Parses the attribute value, checking for W3C compliance, and returns a list with the following tuples: (value, modification tuple) with value a multiplication factor.
nl::ssml_apply_intonation_changes	utt	Applies tags.

Continued on next page

Table B.4 – continued from previous page

Function name	Arguments	Comment
nl::ssml_change_range	utt range	Changes the parameter 'N to range. range is a tuple containing ("absolute",value), ("relative",value) or ("shift",value). Saves the old value of 'N to be restored before the next word is processed.
nl::ssml_restore_f0range	none	Restores the f0 range parameter from backup.
nl::ssml_apply_range_changes	(utt target)	Is called for every word in R:Word before intonation is applied. Checks for the attribute and Parameter.Sets 'N (NeXTeNS uses this parameter for F0 range) accordingly. target is the current target in R:ToneTargets.
ssml_parse_volume_attribute	attribute	Checks attribute for W3C compliance and returns a linear multiplication factor.
ssml_apply_volume_changes	utt	Scans the SSMLTree for and rescales the entire utterance accordingly.
ssml_after_analysis_debug	utt	Your favourite debugcode here.

B.4 SSML Variable list

Variable name	Comment
W3C_NUMBER	A regular expression to verify if a number is conform the W3C specifications.
W3C_SSML_PITCH_VALUE	A regular expression to verify if a argument value is conform the W3C specifications.
W3C_SSML_CONTOUR_VALUE	A regular expression to verify if a argument value is conform the W3C specifications.
W3C_SSML_VOLUME_VALUE	A regular expression to verify if a argument value is conform the W3C specifications.
ssml_pitch_base_map	The lookup table to translate x-high,high,medium,default,low,x-low into appropriate pitch multiplication factors.
ssml_pitch_range_map	The lookup table to translate x-high,high,medium,default,low,x-low into appropriate pitch-range multiplication factors.
ssml_rate_speed_map	The lookup table to translate x-fast,fast,medium,default,slow,x-slow into appropriate speechrate multiplication factors.
ssml_volume_level_map	The lookup table to translate x-loud,loud,default,medium,soft,x-soft,silent into appropriate volume percentages.
nl::ssml_pause_map	The lookup table to translate x-strong,strong,medium,weak,x-weak,none into appropriat break-duration values (ms).
ssml_elements	The big list of ssml tags and what should be done when each is encountered.

Table B.5: SSML Variables