Swedish text-to-speech
in
Festival Speech Synthesis System

Maria Skeppstedt

March 28, 2002
Swedish text-to-speech in Festival Speech Synthesis System
by Maria Skeppstedt

Master’s Thesis in Speech Communication (20 credits)
at the School of Computer Science and Engineering,
Royal Institute of Technology
Supervisor at Voxi AB was Nathalie Talbot
Supervisor at Speech Music and Hearing was Björn Granström
Examiner was Björn Granström

Svenskt text till tal med Festival Speech Synthesis System
av Maria Skeppstedt

Examensarbete i talkommunikation om 20 poäng
vid Programmet för Datateknik,
Kungliga Tekniska Högskolan
Handledare på Voxi AB var Nathalie Talbot
Handledare på Tal Musik och Hörsel var Björn Granström
Examinator var Björn Granström
Abstract

The aim of this master thesis is to determine if it is possible to construct Swedish text-to-speech that results in a comprehensible diphone voice, by using only simple methods provided by Festival Speech Synthesis System and the most basic recording equipment. In order to find this out, a synthetic voice was constructed and evaluated.

To construct the voice, it was firstly established which diphones that should be included in a Swedish diphone database. Most of these diphones were thereafter recorded, but some, for example diphones consisting of vowel-to-vowel transitions were excluded. The diphone database was recorded on a portable computer, using the voice of an untrained speaker.

After the recording, approximate rules for grapheme-to-phoneme conversion and prosody were constructed and implemented.

Finally, the constructed voice was evaluated. The conclusion of this evaluation was that the voice can not be characterized as comprehensible. It was also concluded that this can not be explained only by the incomplete diphone database, or the shortcomings of the grapheme-to-phoneme conversion. It was, however, also determined that the voice is sometimes possible to comprehend, and that it therefore is not unlikely that small improvements in each module would give rise to a comprehensible voice.
Sammanfattning

Detta är examensarbete har till syfte att undersöka om det är möjligt att konstruera svenskt text-till-tal som resulterar i en förståelig difonröst, genom att endast använda enkla metoder inom Festival Speech Synthesis System och enkel inspelningsutrustning. För att ta reda på detta konstruerade och utvärderades en syntetisk röst.

För konstruera den syntetiska rösten, undersöktes först vilka difoner som en svensk difondatabas bör innehålla. De flesta av dessa difoner spelades sedan in, men några, till exempel difoner bestående av övergången mellan två vokaler, togs inte med. Difondatabasen spelades in på en bärbar dator, med en otränad talare.

Efter det att inspelnningen hade gjorts, konstruerades och implementerades approximativa regler för omvandling från grafem till fonem och för prosodi.

Till sist utvärderades den konstruerade rösten. Utifrån denna utvärdering kunde slutsatsen dras att rösten i många fall inte var möjlig att förstå. Det kunde också konstateras att detta inte enbart kunde förklaras med den ofullständiga difondatabasen eller med brister i transskriberingsreglerna. Det kunde emellertid också fastställas att, eftersom den konstruerade rösten ibland var förståelig, så är det inte osannolikt att små förbättringar i varje modul skulle ge upphov till en röst som är möjlig att förstå.
Acknowledgements

Jag vill tacka följande personer:

Contents

0.1 Background ............................................................... 1
0.2 Aim ........................................................................... 1
0.3 Method ....................................................................... 2
0.4 Structure of the report ................................................. 2

I Implementation ............................................................... 3

1 Introduction to Festival Speech Synthesis System ............... 5
  1.1 Voices previously constructed in Festival ................. 6
     1.1.1 Swedish voices constructed in Festival ........... 6

2 Text-to-speech in Festival ............................................... 7
  2.1 Text-to-speech systems .......................................... 7
  2.2 The modules in Festival .......................................... 7
     2.2.1 The natural language processing module .......... 7
     2.2.2 The digital speech processing module ............ 8

3 Diphone synthesis .......................................................... 9
  3.1 Why diphones are used as concatenation segments .... 9
  3.2 Concatenating and modifying the diphones ............... 10

4 Constructing a diphone database ...................................... 13
  4.1 Which allophones to include .................................... 15
     4.1.1 Pronunciation of /j/ ......................................... 16
     4.1.2 Pronunciation of /r/ ......................................... 16
     4.1.3 Consonant combinations that give rise to retroflexes . 16
     4.1.4 Aspirated/Unaspirated plosives ....................... 17
     4.1.5 Voiced/unvoiced plosives ............................... 18
     4.1.6 Nasals ......................................................... 18
     4.1.7 Voiced/unvoiced /v/ ....................................... 19
     4.1.8 Remaining consonant phonemes .................... 19
4.1.9 Vowels .................................................. 19
4.1.10 Pronunciation of /e/ and /e/ ......................... 20
4.1.11 Pronunciation of the letters ò and ä before retroflex consonants 20
4.1.12 Conclusion ............................................. 20
4.2 Which combinations of phonemes to record ................. 21
4.2.1 Phone-to-phone transitions that do not exist in Swedish ... 21
4.2.2 Other diphones that were excluded ..................... 22
4.2.3 Conclusion ............................................. 23
4.3 Constructing carrier words for the recording of the diphones .... 23
4.4 Recording the database ................................... 24

5 Grapheme-to-phoneme conversion ................................ 27
5.1 Determining morpheme boundaries .......................... 27
5.1.1 Determine boundaries between root morphemes and inflectional morphemes .......................... 28
5.1.2 Determining boundaries between root morphemes ........ 29
5.2 Determining word stress ................................... 33
5.2.1 Lexical stress versus phrase stress ..................... 33
5.2.2 Rules for stress ....................................... 33
5.2.3 The implementation of detecting the stressable syllables .. 36
5.3 Determining if a word has accent I or accent II ............. 37
5.3.1 Swedish word accent .................................... 37
5.3.2 Rules for determining if a word has accent I or accent II .. 37
5.3.3 The implementation of determining if a word has accent I or accent II ............................................. 38
5.4 Determining if a vowel is long or short ...................... 39
5.4.1 Rules for whether a vowel is short or long ................ 39
5.4.2 Which of the rules for short and long vowels that were implemented .............................. 40
5.4.3 The implementation of the rules for short or long vowel ... 41
5.5 The actual letter to sound conversion ......................... 41
5.6 Syllabification ............................................. 42
5.7 Concatenating the subwords and adjusting the accent .......... 42
5.8 Post-lexical processing ..................................... 43

6 Prosody .................................................. 45
6.1 Phrasing .................................................. 45
6.2 Pauses in speech ......................................... 46
6.3 F0 modelling .............................................. 47
6.3.1 The pitch contour of accent I and accent II ............. 47
6.3.2 The implementation of the pitch contour .................. 48
6.3.3 The pitch contour on a phrase level .......................... 49
6.4 Duration .................................................................. 50
  6.4.1 Rules for duration .............................................. 50
  6.4.2 The implementation of duration ......................... 50

II Evaluation .................................................................. 51

7 Methods for evaluation of text-to-speech systems ........ 53

8 The evaluation ............................................................ 55
  8.1 The first test .......................................................... 55
    8.1.1 Results ......................................................... 56
    8.1.2 Discussion .................................................. 57
  8.2 A subjective assessment ......................................... 57
    8.2.1 Results ......................................................... 57
    8.2.2 Discussion .................................................. 57
  8.3 The second test .................................................... 58
    8.3.1 Results ......................................................... 59
    8.3.2 Discussion .................................................. 59
  8.4 Comments on the constructed voice ....................... 60
    8.4.1 Comments made by the test subjects .................. 60
    8.4.2 My subjective assessment ............................... 60

9 Conclusion .................................................................. 63
Introduction

0.1 Background

Festival Speech Synthesis System, which is being developed at the University of Edinburgh, offers a framework that enables the construction of new synthetic voices and text-to-speech synthesis for these.

One of the ultimate goals for the development of Festival is to construct a system in which anyone, without much effort, would be able to build his or her own high quality, natural sounding synthetic voice. Optimally, nothing else than a computer with a microphone and Festival Speech Synthesis System installed on it should be needed. Except, of course, a person that is prepared to spend a few hours reading words into the microphone.[3]

Naturally, this is not possible today. But how far away from this goal is the Festival project? Does an attempt to construct Festival text-to-speech by simple methods result in anything at all? Does it result in a synthetic voice with output that can be comprehended, or in a voice which outputs unintelligible sounds?

0.2 Aim

The purpose of this thesis is to investigate whether it is possible to build Swedish text-to-speech that results in a comprehensible diphone synthesis voice, through simple methods provided by Festival Speech Synthesis System and by using the most basic equipment. 'Simple methods' refers to approximate rule-based approaches for determining prosody and for converting from text to phonetic transcription. 'Basic equipment' refers to accomplishing the recording of the diphone database on a portable computer, using my own voice, thus the voice of an untrained speaker.
0.3 Method

The method consists of two parts: Firstly the voice was constructed with the methods and equipment described above, and thereafter the constructed voice was evaluated.

0.4 Structure of the report

The structure of the report corresponds to the two parts of the method. The first part describes the implementation of the text-to-speech synthesis, and the second part describes the evaluation. Each section in the report starts with introductory theory, and then the actual implementation or realization is described.
Part I

Implementation
Chapter 1

Introduction to Festival Speech Synthesis System

Festival Speech Synthesis System, which from now on will be referred to as Festival, is developed at CSTR\textsuperscript{1}, University of Edinburgh by, among others, Alan W Black, Richard Caley and Paul Taylor. The development started in April 1996, and the system is in constant development.

Festival offers a framework for building text-to-speech systems. The system is written in C++, and programs included in the Edinburgh Speech Tools are utilized for the construction of diphone synthesis, unit selection synthesis or limited domain synthesis. A Scheme based scripting language is also included and a Scheme interpreter is used for specifying parameters and flow of control. This means that parameters can be altered at run time, without any need for a re-compiling of the system. This scripting language is utilized for constructing procedures and defining parameters when building a new voice.

Festival does not only aim at users which are interested in constructing a new synthetic voice, but also at users which need to embed one of the existing Festival voices within an application. It also aims at users which need a framework for speech synthesis research, for example in order to evaluate or test one aspect of text-to-speech. The part of text-to-speech that is to be studied can be incorporated into a module and combined with the framework of existing modules in the Festival system. This gives an opportunity to concentrate on the research topic, instead of on building a text-to-speech system.

\textsuperscript{1}The Centre for Speech Technology Research
1.1 Voices previously constructed in Festival

The first effort to build a Festival voice, a Scots Gaelic synthesizer, was made in the summer of 1996, a few months after the development of Festival had started. Since the processes involved in building Festival voices were not as simple as they are today, the project was never finished. It did, however, show that it was possible to construct a voice, and later that summer the constructors of Festival built a Spanish synthesizer using Festival. The following year a Greek and a Basque synthesizer were built, and not long thereafter voices in many other languages were built.[3]

1.1.1 Swedish voices constructed in Festival

In spring 1999 students at the computational linguistics programme at Göteborg University tried to make the first Swedish Festival voice. This project was, however, not finished.[10]

Later in 1999 a Swedish Festival voice, using the Lukas diphone database in MBROLA format, was constructed by Johan Frid at Lund University. He describes the voice as follows:

"This voice provides a Swedish male voice. It uses letter to sound rules for pronunciations. Prosodic phrasing is provided by a simple CART tree. Intonation is provided by the letter to sound rules by assigning stress and accent to every long vowel (it's really just a test model) and an F0 contour is generated by introducing a hat pattern at every accented vowel. The durations are given from a simple list of durations for each phoneme."[11]

In spring 2000 a Swedish unit selection speech synthesis² was constructed by Jessica Granberg and Hanna Lindgren, as a master thesis project in computational linguistics at Stockholm University. The aim of the project was to explore the possibilities of unit selection, and the resulting synthesized speech had a varying quality, sometimes producing natural-sounding high-quality speech and sometimes producing unintelligible speech.[12]

---

²See section 2.2.2
Chapter 2

Text-to-speech in Festival

2.1 Text-to-speech systems

The definition of a text-to-speech system is a computer-based system with the ability to read any text aloud. A text-to-speech system consists of two main parts: The first part is the natural language processing module, which produces a phonetic transcription of the input text, together with the prosodic information, i.e. information about e.g. pitch, duration and phrasing. The second part is the digital signal-processing module, which transforms the information given by the natural language processing model into synthesized speech.[14]

2.2 The modules in Festival

2.2.1 The natural language processing module

The natural language processing module included in Festival is built up of the following components:

- A text analysis module, in which tokens and numbers are translated into text.

- A grapheme-to-phoneme module, in which each word is first transcribed separately, either using a lexicon or using some rule-based or statistical method. Thereafter, post-lexical rules are applied to modify the transcription of the words according to coarticulation effects that reach over word boundaries.

- A phrasing module, in which the text is segmented into phrases.

- A pitch module, in which the tone curve of the utterances is determined.
• A duration module, in which the duration of the segments in the utterance is determined.

2.2.2 The digital speech processing module

The digital signal-processing module included in Festival uses concatenative methods for generating waveforms. That is, segments selected from recorded human speech are modified by a signal processing function and concatenated into synthesized speech.\(^1\) The two concatenative methods that are included in Festival are unit selection synthesis and diphone synthesis. For unit selection synthesis, a large database of natural speech, which contains multiple occurrences of the same speech segments, but pronounced in different prosodic contexts, is needed. When synthesising the speech, the most appropriate segments of the most appropriate lengths, based on prosodic, phonetic and acoustic features, are extracted from the speech database, and concatenated into speech. Since the synthesising strategy in this study is diphone synthesis, it is described in more detail in the following chapter.

\(^1\)The most common other synthesising methods, which are not included in Festival, are formant synthesis and articulatory synthesis.
Chapter 3

Diphone synthesis

As opposed to unit selection synthesis, which is a relatively new concatenative strategy, the first diphone system was demonstrated already in 1967, at the MIT: Conference on Speech Communication and Processing.[15]

In diphone synthesis, speech segments named diphones are concatenated into a desired utterance. A diphone is a speech unit, that starts in the middle of one phone and ends in the middle of the following phone. Each diphone thus consists of a pair, in which one is the second half of a phone and the other is the first half of a phone. Consequently, when concatenating the diphones into speech, a diphone ending with a first half of one phone is always followed by a diphone starting with a second half of that phone. For example, the following diphones are concatenated in order to produce the word meet: 'silence + m', 'm + i'; 'i: + t' and 't + silence'. Thus, the first one of these diphones consists of 'silence + the first half of an m' and the second one consists of 'the second half of an m + the first half of an i:'. The third diphone consists of 'the second half of an i:+ the first half of a t' and the last one of 'the second half of a t + silence'. The concatenation takes place in the middle of the phones since it, in general, is the most stable part of the speech. That is, the change in articulation is at its minimum at the middle of the phones.[15]

3.1 Why diphones are used as concatenation segments

The reason why a combination of two phones are extracted and utilized for concatenation, instead of concatenating speech segments consisting of only one phone, is the coarticulation effects of natural speech. That is, the fact that the pronunciation of a phone is affected by other phones positioned close to it. The /s/ in sweet, for example, is produced with rounded lips, since it is affected by the ar-
articulation of the following phone. Concatenating speech units consisting of only one phone would thus yield synthesized speech with a pronunciation very far from the natural one. When instead using diphones, the coarticulation on the first half of a phone caused by a preceding phone and the coarticulation on the second half of the phone caused by the following phone is incorporated. Therefore, diphone synthesis gives rise to a more natural pronunciation.

However, in natural speech coarticulation often affects a whole phone, rather than just parts of it, and coarticulation effects can also reach further than to the two surrounding phones. For instance, in some languages an assimilation effect devoices phones positioned next to unvoiced segments. Partial devoicing is approximately covered by the use of diphones, since the half of the phone positioned next to the unvoiced segment is unvoiced, whereas total devoicing is not. Covering such cases requires that the allophones arising from the coarticular effects are included in the diphone database. Such a strategy, where several allophones are included in the database is denoted allo-diphone synthesis. Longer speech units, as for example triphones or disyllables, which cover more coarticular effects can also be used as concatenation segments.

The advantage of using diphones rather than longer speech units is, however, that the size of the speech segment database is comparatively small.[13]

### 3.2 Concatenating and modifying the diphones

The speech from which the diphones are extracted should optimally be produced with a constant pitch, lying in the middle of the range of the speaker. However, in natural speech pitch and duration vary, for instance due to different stress levels or different positions in the utterance. Unless the synthesized voice is to be produced in a monotone, the pitch and duration of a diphone need to be modified to a level appropriate to the context into which it is being concatenated. In order to obtain these prosodic variations with a minimal distortion to the speech signal there are various techniques. Most of these methods are pitch synchronous, which means that the speech is partitioned into short signals consisting of a single pitch period. In accordance with figure 3.1, a signal starts at the onset of a pitch pulse and ends just before the onset of the following pulse in voiced parts of speech. In the unvoiced parts some arbitrary section is chosen.

To shorten the duration of a diphone, some of the pitch pulses are deleted and to lengthen the duration some pitch pulses are duplicated. A change in duration by a factor of 0.5 to 2 should be possible without affecting the quality too much.

To increase the pitch, the signals are moved to overlap and to lower the pitch the signals are moved further apart. Similarly, it should be possible to apply a factor of 0.5 to 2 on the pitch of the segments without affecting the position and bandwidth.
of the formants, that is without obtaining a distorted voice. Using diphones instead of longer speech units leads to a high density of coarticulation points, one per phone, which requires efficient and high quality concatenation.[4][13]

Supported by Festival are PSOLA (Pitch Synchronous Overlap Add) technique, MBROLA technique and RELP (residual excited LPC), but only the RELP technique, which uses an LPC analysis and reconstruction method for modifying the pitch, is included in the free version of Festival.[4]
Chapter 4

Constructing a diphone database

The first step in building a new diphone voice in Festival is to record a diphone database. In order to determine which diphones to record, a list of all possible phone-to-phone transitions in the language must be established. To obtain this list the following must be analysed: Firstly, which phonemes the language consists of. Thereafter, if there are any allophones of these phonemes that are salient enough to give rise to additional diphones, and if there are any consonant clusters that are pronounced in such a way that diphones consisting of these consonant clusters ought to be recorded. Finally, it must be determined into which diphones the phonemes, their allophones and the consonant clusters can be combined, since it is not certain that all the transitions between the phones of the language exist. Unnecessary diphones should not be included in the diphone database, since when the number of diphones that are recorded increases, so does also the risk that the recording will be of low quality. The reason for this is that long recording sessions increase the amount of pronunciation errors and the need for breaks, and decrease the quality of the speech produced by the speaker.

When it has been decided which diphones to record it must also be determined how to record them. One technique for articulating the diphones properly is to record them in the context of carrier words. These carrier words could either be real or nonsense words, which contain the diphones that are to be recorded. The advantage of using real words is that the articulation of the diphones could be more natural than when using nonsense words. The disadvantages are, however, that it takes a lot of effort finding natural examples that contain the desired diphones and that the labelling of the diphones is much easier when using nonsense words. Moreover, when recording a diphone database, the phones ought to be pronounced with a constant duration and pitch. Constant pitch and duration might for most speakers be easier to achieve with nonsense words than with natural words, since the speaker is normally used to vary the pitch and duration when pronouncing the natural words. If possible, it is often most appropriate to position the diphone
that is to be recorded in the middle of the carrier word. The reason for this is that it is usually more difficult to keep a constant pitch and duration in the beginning and in the end of an utterance.

In order to articulate the diphones properly, a synthesized voice, which reads the nonsense words with a constant pitch and duration, can be used. If the speaker mimics this prompt when recording the diphones, it is easier to produce the carrier words with the desired constant pitch and duration. Also, when mimicking another voice, the risk that the speaker articulates the wrong diphone decreases. The pitch of the prompt voice ought to be set to the middle of the range for the speaker, in order to facilitate the mimicking. If there already is a synthetic voice in the language for which the diphone database is to be recorded, this voice can be used for synthesising the prompts. If no such voice exists and there is a voice of another language, which phone set is close to the set of the new language, then this voice may be utilized as an approximation of the correct articulation. The disadvantage of using phones from another language is, however, that they could influence the speaker to pronounce the carrier words with the same foreign accent as the prompt voice.

There is a Festival script which, given a list of carrier words, synthesizes prompts using the voice that is to be mimicked. If the database is to be recorded using a computer, there is also a Festival script that can be utilized for this recording. This script, which uses the same list of carrier words, firstly plays the prompt and prints the transcription of the carrier word on the screen. Thereafter it records the speaker. If nonsense words are to be used as carrier words, then these can also be generated using a Festival script, which evokes Scheme procedures that constructs the desired carrier words.

Before these Festival scripts for constructing the prompt words and for recording can be run, Scheme procedures that generate carrier words containing the diphones must be written. To be able to write these Scheme procedures, it must thus firstly be decided which carrier words that are most appropriate for the different types of diphones. One type of carrier word could, for example, suit consonant-to-consonant diphones, another type consonant-to-vowel, a third type could be appropriate for some kinds of allophones and so on.

Recording directly to a computer could decrease the quality of the diphone database. If the recording takes place in the same room as the computer, it could add noise to the recorded voice, and if the quality of the computer audio hardware is not high enough it might also give a bad result. The advantage of using a computer is, however, that the Festival script for recording can be utilized, which automatically reads the prompt voice, records the speaker and, most importantly, stores each carrier word in a separate file. When not recording directly to a computer, the recorded carrier words have to be transferred from the tapes to the computer and moreover be split up into separate files.
The recording should, optimally, take place in an echoic chamber, but general recording studios are also possible. A head-mounted microphone can be used to ensure that the distance between the speaker and the microphone is kept constant. Optimally, in order to be able to keep the recorded voice as constant as possible, the recording ought to be completed in one session. This might, however, not be possible if a large diphone database is to be recorded. Even if the recording is not split up into several recording sessions, the recording conditions and environment should be repeatable. The speaker should thus not have a cold or something similar that might affect the voice and that can not be repeated for the next recording session or for re-recordings of diphones that were mispronounced.

After the recording of the carrier words has been done, the recorded files must be labelled. By labelling the recorded files, it is meant marking out the start and end of the phones in the carrier word. The labelling can either be done by hand or by an automatic labelling program which is included in Festival. The automatic labelling matches the recorded carrier words with the words that were synthesized as prompt words, and can thereby determine the location of the phones in the sound file.

The constructed labelling files are utilized by a Festival script to build a diphone index, which is a file that lists in which sound file each diphone can be found, along with its start time, midpoint and end time. This diphone index is utilized for locating the diphones when synthesising the voice. The diphones are then extracted from the middle of its first phone to the middle of its second phone.[3]

### 4.1 Which phonemes and allophones to include in the database

The first step in constructing a diphone database was thus to determine which phonemes that were to be included in the Swedish diphone database, and whether there were any allophones of these phonemes that also should be contained in the database.

Since the synthetic voice was to resemble my own natural voice, the phonemes which I normally use, and also the most salient allophones of these phonemes in my dialect, had to be included in the database. My dialect has been influenced by my parents' dialects¹, the way people speak in Stockholm, and by some kind of 'standard Swedish'.

In order to reduce the size of the diphone database, neither all possible allophones of the chosen phonemes, nor all possible phonemes, were recorded. In addition to consulting phonetic literature for deciding which phonemes to include,

¹My mother originates from Sundsvall and my father from Gnesta
a small test diphone database, without any allophones at all, was recorded to make it possible to determine the effect of excluding allophones. The following is an analysis of some of the phonemes and allophones of my dialect, and motives are also given why some phonemes and allophones were chosen to be included in the database and some were not.

### 4.1.1 Pronunciation of /ʃ/

In the middle and northern parts of Sweden /ʃ/ is pronounced as a voiceless retroflex fricative sound, [ʃ]. In Stockholm /ʃ/ can either be pronounced as a voiceless velar fricative [ʃ], or as [ʃ], depending on the sociolect and gender of the speaker, and on the position of /ʃ/ in the word. In the end position of a word [ʃ] is more usual and in the word initial position [ʃ] is preferred. I normally alter between the retroflex and velar /ʃ/, though I mostly use [ʃ] when it occurs in the word initial position and [ʃ] when it occurs in another position. In the diphone database, however, only the retroflex variant has been included. In the resulting diphone voice the retroflex /ʃ/ was thus used in all positions of a word.[5]

### 4.1.2 Pronunciation of /r/

The most common pronunciation of /r/ in Finland and in the north and middle parts of Sweden is a retroflex or alveolar trill. However, in these dialects it is also common to mix this with other pronunciations, in which the /r/ is pronounced as an approximant or a fricative. It is especially common when the /r/ is in the final position of a word, for example in the word pojkar or when it is positioned in the middle of two vowels, for example in word bara. This approximant or fricative pronunciation of /r/ does, however, occur in many positions and is not only dependent on the phonetic context.

My pronunciation of /r/ alters between the different pronunciations. When recording the diphone database, however, I tried to produce the /r/ with a uniform pronunciation, probably closest to an alveolar fricative.[5]

### 4.1.3 Consonant combinations that give rise to retroflexes

When /r/, in my dialect, is followed by either /t/, /d/, /s/ or /n/, the two consonants form a combination which is pronounced as one retroflex phone. The last phone in each of the words kort, bord, fors and barn are thus pronounced as /t/, /d/, /n/ and /s/, respectively. The retroflex phonemes also appear between word boundaries, thus when the first of two adjacent words ends with an /r/ and the second starts with either /t/, /d/, /s/ or /n/. The combination of /r/ followed
by an /l/ gives, in some dialects, also rise to a retroflex phone. This retroflex is, however, not as usual as the other four, and I do not use it.

In the first test diphone database, not all of these retroflex consonants were included, which resulted in a very unnatural pronunciation. When recording the real database, all the usual retroflex consonants were therefore included. The diphone combinations of the phones /r/ and /t/, /l/, /s/ or /n/, not pronounced as one retroflex but as two separate consonants, were also included. The reason for this is that these combinations do not result in a retroflex combination when the /r/ is produced with long duration. A consonant with long duration appears directly after a stressed short vowel. In the words barrträd and kurd, for example, the /r/ has a long duration and therefore the /r/ and the following consonant are pronounced as two separate phones.[5][6]

4.1.4 Aspirated/Unaspirated plosives

The plosives /p/, /t/, /k/ and /t/ are aspirated under the following circumstances:

- When they are positioned after a pause or as the first phone in a morpheme, as in the words par, tå, kår, fotpall and fartyg.

- When they appear as the first phone in a stressed syllable, even if this syllable is not the start of a morpheme, except when the plosives are preceded by an /s/. In the words tapir, patē, förtal, raket, /p/, /t/, /rt/ and /k/ are thus aspirated.

- When they occur as the last phone before a pause, for example in the words lapp, fatt, hår and lakk.

The plosives /p/, /t/, /k/ and /t/ are, on the other hand, unaspirated under the following circumstances:

- When they appear after an /s/ in the same morpheme, as in the words spar, stå, förstå, skåra, estet, asket.

- When they are positioned in front of a voiceless segment, as in the words raps, bets, harts and makt.

- When they are followed by an unstressed syllable, except when that syllable is the first in a morpheme. In the words lappen, fatta, tårta and lakkens /p/, /t/, /k/ and /t/ are thus unaspirated.[6]

In order to reduce the size of the diphone database, no unaspirated allophones of the plosives /p/, /t/, /k/ and /t/ were specially recorded. However, when
recording the diphones consisting of these plosives, the plosives were sometimes pronounced with an aspiration and sometimes without, depending on the carrier word and on which phone the plosive was combined with. This gives rise to a synthesized voice which sometimes produces an aspirated plosive, where it should have been an unaspirated one.

4.1.5 Voiced/unvoiced plosives

The plosives /b/, /d/, /g/ and /d/, which normally are voiced, become under some circumstances devoiced. They remain voiced when they are positioned between voiced segments, for example in labben, fadda, vårda and en gång. They are, however, only partly voiced or unvoiced when they are positioned in front of or after an unvoiced segment, as in the words snabbt, sagt and resdag, and when they are positioned before or after a pause. When these plosives are positioned next to voiceless segments they might be impossible to distinguish from their corresponding unvoiced plosive. The pairs potatisblast/potatisplast and förstä/forstå are examples of this. However, the plosives /b/, /d/, /g/ and /d/ are always unaspirated, whereas the corresponding unvoiced plosives in some cases are aspirated.[6]

In order to save time, no diphones containing the unvoiced allophones of the plosives were, however, added to the diphone database.

4.1.6 Nasals

The nasals /m/, /n/ and /ŋ/ which are normally voiced, become devoiced when they, in addition to following an unvoiced segment occurs in the final position of a word. They are also devoiced when preceded by an aspirated plosive. The words rytm and knapp are examples of this.[6]

No unvoiced nasals were explicitly added to the database, and when recording the diphones some nasals were thus produced as voiced and others as unvoiced, depending on which phone they were combined with. However, in the first test diphone database, the lack of separate diphones with unvoiced nasals was not noticeable.

Another allophonic variation that takes place when producing nasals is nasal assimilation. This means that the nasal is articulated at the same position as the consonant following or preceding it. The /n/ in inbilla, for example, is pronounced as an /m/. The dental nasal /n/ is thus affected by the bilabial /b/ that follows it and is therefore instead pronounced as a bilabial nasal, which is an /m/. It is, however, also possible to pronounce the nasals without nasal assimilation, which does not sound incorrect, only well articulated.[6]
Since it is not incorrect to pronounce the nasals without nasal assimilation, no additional diphones consisting of the assimilated nasals were added to the diphone database. The nasals were instead recorded without assimilation with the other consonant in the diphone. However, in cases in which the nasal assimilation resulted in diphones that were already contained in the database, the grapheme-to-phone rules were utilized for transcribing the nasals into the assimilated form. Thus, the n in imbillia was transcribed as an /m/ since diphones consisting of /m/ were contained in the database. Accordingly, n preceding p was transcribed as /m/ and n preceding k was transcribed as /ŋ/.

4.1.7 Voiced/unvoiced /v/

The voiced phoneme /v/ is voiced when it is positioned between other voiced segments, and unvoiced or partly voiced when it is positioned before or after an unvoiced segment.[6]

To reduce the size of the diphone database, however, no unvoiced allophone of /v/ was added to the diphone database.

4.1.8 Remaining consonant phonemes

The phonemes /j/, /l/, /ʃ/, /h/, /s/ and /ç/ are other Swedish consonant phonemes, which show more or less allophonic variation. The normally voiced phoneme /l/ might be devoiced when positioned next to aspirated plosives, and the unvoiced phoneme /h/ might become voiced when positioned between two voiced segments. Also the phoneme /j/ might show allophonic variation depending on whether it is positioned next to voiced or unvoiced segments.[6] However, none of these variations were considered significant enough for adding any additional allophones to the database.

4.1.9 Vowels

In Swedish there are two main versions of a vowel, one long and one short. A short vowel has a shorter duration than the corresponding long vowel, but there are also other pronunciation differences between them.[6]

Both the long and the short version of the each vowel were added to the diphone database, with one exception which is described in the following section. In some positions the letter e is pronounced as /œ/. Also, all vowels, when they occur in a very reduced position, can be pronounced as schwa. Therefore /œ/ was also added to the database.

In a stressed syllable, both long and short vowels can occur, but only short vowels occur in an unstressed syllable. However, in some unstressed positions
there are speakers who pronounce the vowels with the quality of a long vowel but with the duration of a short vowel.[6] Since this only happens in a few positions and only for some speakers, it was not accounted for when constructing the diphone database.

4.1.10 Pronunciation of /ɛ/ and /e/

Most Swedish dialects contain both the short phoneme /ɛ/ and the short phoneme /e/. That is, there is a difference between how the letters å and e are pronounced, for example in the words teckning and tåckning. In the dialects spoken in the eastern parts of Sweden, however, there is no such pronunciation difference between those two short phonemes. The letters å and e in words as teckning and tåckning are instead both pronounced as /ɛ/. Since my dialect is one of these that are spoken in eastern Sweden, only the phoneme /ɛ/ was added to the database.[5]

4.1.11 Pronunciation of the letters ö and ä before retroflex consonants

In many Swedish dialects there is a difference between how the letters ö and ä are normally pronounced and how they are pronounced when they, within the same morpheme, are followed by /r/, /t/, /s/, /d/ and /n/. They are normally pronounced as [e] and [œ]. However, when positioned before /r/, /t/, /s/, /d/ and /n/ they are more open, thus pronounced as /æ/ and /œ/

This allophonic variation of the pronunciation of /ɛ/ and /œ/ are hardly noticeable in some dialects, but salient in others, in which my Stockholm dialect is included. Concatenating a diphone which contains the normal, more closed variants of the more open allophones of /ɛ/ and /œ/ might thus sound unnatural. Therefore, also the more open allophones of /ɛ/ and /œ/ have been included in the diphone database. This variation in pronunciation concerns both long and short /ɛ/ and /œ/. It gives consequently rise to a total of eight allophones of /ɛ/ and /œ/.[5]

4.1.12 Conclusion

At least the following phonemes and allophones need thus to be recorded in order to obtain a synthesized voice using my dialect:

- Long and short vowels: /i:/ /i/ /e:/ /e/ /æ:/ /æ/ /y:/ /y/ /œ:/ /œ/ /u:/
/ʊ/ /u:/ /o:/ /œ:/ /ɔ/ /œ:/ /a/

20
• Vowel allophones: /æ/ /ə/ /ɶ/ /ɶ/ 
• Consonants: /p/ /b/ /t/ /d/ /k/ /ɡ/ /f/ /v/ /s/ /h/ /z/ /m/ /n/ /ŋ/ /r/ /l/ /j/ /u/ /ɥ/ /ʂ/ 

4.2 Determining which combinations of phonemes to record

If all the phonemes and their allophones that are listed above would be combined into diphones, it would give rise to 1849 diphones, i.e. the number of phones squared. The diphone combinations that are impossible in Swedish were, however, not recorded. In order to save time through further reducing the size of the diphone database, some diphone combinations that are possible in Swedish were also removed.

4.2.1 Phone-to-phone transitions that do not exist in Swedish

The diphones consisting of phone-to-phone transitions, which do not exist in Swedish are in most cases not needed for synthesizing a Swedish voice and are moreover difficult for the speaker to pronounce. It was therefore needed to be determined which phone-to-phone transitions that are impossible in Swedish, in order to exclude these from the list of diphones that were to be recorded.

There are phonotactic rules for which phonemes that can follow a given phoneme, and which that can not. For example, an /f/ can not, within the same word, be followed by a /b/. It can, however, be followed by an /r/, as in the word frukost.[7] These rules are, however, not enough for determining which diphones to exclude, since in fluent speech there is normally not a distinct pause between words. When synthesizing the words tuff basker, for example, no pause would normally be made between tuff and basker and the diphone /f+b/ would thus be needed. Swedish words can also be combined into compounds. The words soffa and bord, for example, can be combined into the compound soffbord, in which the diphone /f+b/ also occurs. However, there are also phonotactic rules dictating which phonemes that do not occur in the front or the final position of a word. These rules can, together with the phonotactic combination rules, be used in order to determine which phone-to-phone transitions that do not exist in Swedish.

---

[7]When synthesizing foreign words, for example names, these phone-to-phone transitions are needed. In addition, some foreign words, e.g. English names, are normally pronounced with English phonemes, and therefore both English phonemes and English phone-to-phone transitions are also needed in a good diphone database. However, since it is only a simplified database, no foreign phone-to-phone transitions or phonemes were included.
The phonemes /h/ and /q/ do normally not occur in the final position of a word, and according to the combination rules, they can not be followed by another consonant. Therefore, the diphone combinations /h+consonant/ and /q+consonant/ do not exist in Swedish, and those diphones were consequently not recorded.

The phonemes /ŋ/, /t/ /d/ and /ŋ/ neither occur in the front position of a word, nor after a consonant inside a word. Therefore, the diphones consisting of the combinations /consonant+ŋ/, /consonant+t/, /consonant+d/ and consonant+ŋ/ were not recorded.

The allophones /æ:/ /æ/ /ɑː:/ /ɑː/ only occur when followed by /r/ /t/ /s/ /d/ and /ŋ/. Therefore, only diphones in which these allophones were followed by /r/ /t/ /s/ /d/ and /ŋ/ were recorded. Neither were the diphones consisting of these vowels and the consonant phoneme /s/ recorded.

4.2.2 Other diphones that were excluded

In order to reduce the amount of work spent on recording, not all phonotactically possible combinations of the phones and allophones that were listed in section 4.1.12 were recorded.

The combination of any of the vowels /ɛ/, /ɛː/, /œː/ or /œ/ followed by /r/, /t/, /d/ and /ŋ/ are only possible if there is a morpheme boundary between the vowel and the consonants. The reason for this is that the vowels are normally changed into: /æː/ /æ/ /œː/ and /œ/ if they precede /r/, /t/, /d/ or /ŋ/ within the same morpheme. However, to reduce the number of recorded diphones, the combination of the vowels /ɛ/, /ɛː/, /œː/ and /œ/ and the consonants /r/, /t/, /d/ or /ŋ/ were not recorded. Neither were the diphones consisting of these vowels and the consonant phoneme /s/ recorded.

The diphones that are a combination of one phone, for example /ɛ+ɛ/ or /r+r/ were also excluded. When synthesising an utterance containing two identical consonants in a row, those two consonants were instead replaced by one occurrence of the consonant. For example the utterance jag har rätt contains two adjacent /r/. When synthesising that utterance those two /r/ were thus replaced by one /r/. When an utterance, containing two identical vowels next to each other, was synthesized, a pause was instead positioned between the vowels.

Also all diphones consisting of /vowel + vowel/ were excluded from the diphone database, even though vowel-to-vowel transitions occur, also within words in Swedish. Consequently, not including them in a diphone database leads to a synthesized voice which does not pronounce words containing /vowel + vowel/ diphones correctly. In the synthesized voice, however, /vowel + vowel/ diphones were replaced with the two diphones /vowel+silence/ and /silence+vowel/. Words as slöare, tjugoett and so on were thus synthesized with a silence of 0.3 seconds between the vowels.
4.2.3 Conclusion

The following phone-to-phone transitions do normally not exist in Swedish and were therefore not recorded:

- /h+consonant/
- /ζ+consonant/
- /consonant+n/ 
- /consonant+t/
- /consonant+d/
- /consonant+n/
- /æ:/ /æ/ /œ:/ or /œ/ + consonant, except for the consonants /r/ /t/ /d/ /n/ /ʃ/

The following phone-to-phone transitions are possible in Swedish, but were nevertheless not recorded:

- Vowel + vowel
- The combinations of /ɛ/, /ɛː/, /œː/ or /œ/ and /r/, /t/, /d/, /n/ or /ʃ/
- Diphone-combinations of one phoneme

All other combinations of the above listed phones and allophones were recorded

4.3 Constructing carrier words for the recording of the diphones

After having decided which diphones to record, it was determined whether real or nonsense words were to be utilized as carrier words. Nonsense words were chosen, since it, as described above, simplifies labelling and probably also pronunciation with constant pitch and since it is easier to produce nonsense words than finding real examples containing the diphones. Each nonsense word consisted of one or two of the diphones that were to be recorded. The following nonsense words were used:

The two possible combinations of a consonant and a vowel, thus consonant+vowel and vowel+consonant, were recorded using one nonsense word, namely /d α:
consonant1 vowel consonant1 αː/. The diphones /r+iː/ and /iː+r/, for example, were recorded using the nonsense word /d αː r iː r αː/.

In order to record the diphones consisting of two consonants, the nonsense word /d αː: consonant1 consonant2 αː: d αː/ was used. The diphone /r+m/ , for example, was recorded using the nonsense word /d αː r m αː: d αː/ and the diphone /m+r/ was recorded using the nonsense word /d αː: m r αː: d αː/.

In order to record the diphone containing one of the more open allophones /æː/ /æ/ /ɛː/ /ɛ/ that precede /r/ /t/ /l/ /ŋ/ /ʂ/ a nonsense word with an /r/ following the vowel was used to obtain the correct allophonic pronunciation of the more open vowels. Therefore, nonsense words of the type /d αː: phone more-open-allophone r αː/ were used, for example /d αː b æː r αː/.

The combinations of the more open vowels /æː/, /æ/, /ɛː/, /ɛ/ and the consonants/r/, /t/, /l/, /ŋ/, /ʂ/, also needed to be recorded. In order to do this the nonsense word /d αː: more-open-allophone retroflex-consonant αː/ was used.

Finally, for recording the diphones containing silence, the following nonsense words were used:

- **vowel + silence**: /d αː: d vowel silence/
- **silence + vowel**: /silence vowel d αː/:
- **consonant + silence**: /d αː: d αː: consonant silence/
- **silence + consonant**: /silence consonant αː: d αː/:

Scheme procedures that produce a list of these nonsense words, containing all the diphones that were to be included in the database, were written. This list was thereafter utilized by the Festival scripts for synthesising prompt words and for the recording. The existing Swedish Festival voice, constructed at the Lund University3 was utilized for synthesising the prompt words. The prompt voice was set to have a constant pitch of 230, which is somewhere in the middle of the range for my voice. The prompt voice was also set to pronounce the phones with a longer duration than what is normally recommended for an English Festival prompt voice. The reason for this is that it was discovered to be difficult pronouncing the long Swedish vowels when mimicking a voice with the recommended short duration.

4.4 Recording the database

The first recording took place in an open room, which yielded a voice of very low quality. Therefore, a new recording in a studio at TMH4 was made. A head-

---

3See section 1.1.1
4The Department of Speech, Music and Hearing at KTH, Stockholm

24
mounted microphone on a portable computer was used, and no larynograph signal was recorded. Since the diphones containing vowel-to-vowel transitions were excluded, the recording could be completed in a few hours in one session. After each nonsense word had been recorded the Festival script played the recorded word, thereby enabling an immediate re-recording of the nonsense word if it sounded mispronounced or if the recording sounded distorted. However, a re-recording was only done when there was a very obvious mispronunciation or when the recording sounded very bad.

After the recording, Festival scripts were utilized for labelling and for constructing a diphone index. However, there was not enough time to make sure that each diphone was correctly pronounced, recorded and labelled. This has the effect that the constructed database probably contains diphones that are mispronounced, recorded with a distortion and mislabelled.

When the diphone index had been built, the Festival scripts were utilized for extracting pitchmarks from the waveforms of the carrier words, and thereafter for building LPC parameters. The script for extracting the pitchmarks was modified, since its default parameters were adapted to a male speaker. The minimum and maximum values for the time between the pitch mark peaks were changed from 0.005 and 0.012 seconds, i.e. 200 and 83 Hz, to 0.003 and 0.006 seconds, i.e. 333 and 167 Hz, respectively.
Chapter 5

Grapheme-to-phoneme conversion

When all the symbols in the text have been transformed into graphemes, they, as well as the rest of the text that is to be synthesized, have to be translated into a phonetic transcription. In most voices constructed by Festival a transcription lexicon is utilized for this translation. For words not contained in the lexicon, letter-to-sound rules, either constructed by hand or automatically, could be used. It might, depending on the language, also be possible only to use the letter to sound rules. There are some basic Scheme procedures included in Festival that can be augmented to suit the conversion to phonemes in the desired language. There is also support in Festival for using the lexicon to automatically build a CART tree, which can be utilized to predict the pronunciation of a word. In Festival there is also support for constructing post-lexical rules, i.e. rules for pronunciation that reaches over the word boundaries.[3]

A rule-based approach, with hand-made rules, was used for the grapheme-to-phoneme conversion. The hand-made rules were implemented to work in the following order: The morpheme boundaries of the word are first predicted, and thereafter where stress is put in the word. After that it is predicted if the word has accent I or accent II. Before the actual conversion into phonemes takes place, it is also predicted whether the vowels are long or short.

The nature of all the constructed and implemented grapheme-to-phoneme rules is very approximate. It is thus not at all claimed that the rules perfectly transcribe Swedish text.

5.1 Determining morpheme boundaries

The first step in the grapheme-to-phoneme conversion is to determine the boundaries between the morphemes of the word. There are three reasons why this is needed for the grapheme-to-phoneme conversion:
The first reason is that whether the vowel is short or long is dependent on the number of consonants that both directly follows the vowel and also belong to the same morpheme as the vowel. The ö in dörr, for example, is pronounced as a short vowel, since it is followed by two consonants. The a in daglôk is, however, pronounced as a long vowel, even though it is followed by two consonants, since only the d belongs to the same morpheme as the a. The same is true for u in ljuslina, it is pronounced as a long vowel, since only the s belongs to the same morpheme as the u.

The second reason is that the stress of a compound word is different from the stress of a non-compound word. In a non-compound there is usually only one stressed syllable, whereas in a compound word there are two stressed syllables.

Also the third reason has to do with stress. Some morpheme endings are stressable, the ending -ik is, for example stressable, and therefore stressed in words like musik. The syllable -ik is also stressed in the words musikten and musikskola, since -ik in both cases is a morpheme ending. In order to determine this, it must, however, first be determined that ten and skola do not belong to the same morpheme as musik.

The detection of morpheme boundaries was implemented to work in the following order: Firstly, if the word has any inflectional morphemes, the boundary between these and the root morphemes is first determined. Thereafter, if the word is a compound, the boundaries between the word components are determined. A list\(^1\) containing 120 000 common Swedish words is utilized both for determining inflectional suffixes and the components of the compound.

5.1.1 Determine boundaries between root morphemes and inflectional morphemes

The detection of inflectional morphemes was implemented to function as follows:

It is first determined whether the word ends with an s, which is a suffix denoting genitive for nouns and passive form for verbs. In order to minimize the risk that the s is not an inflectional suffix but the ending of a root morpheme, it is secured that the part of the word preceding the s is contained in the word list. For example, the word plats ends with an s, but since the remaining word plat is not contained in the word list, the s is not marked as a suffix. The word bokens, on the other hand, ends with an s and boken is contained in the word list, therefore the s will be marked as an inflectional suffix.

The next step\(^2\), which determines if the word has any inflectional suffixes, other

---

\(^1\)This word list is a part of the Nordic Words project, and can be found at www.lysator.liu.se/runeberg/words/.

\(^2\)This solution was inspired by the spelling correction tool, Stava.[18]
than $s$, utilizes the inflectional class of the word. That is, the categorisation of the
word according to which suffixes it can have. The words $\text{stol}$ and $\text{kopp}$ belong,
for example, to the same category, since they can both have the suffixes $\text{en ar arna}$, yielding the words $\text{stolar}$, $\text{koppar}$, $\text{stolen}$, $\text{koppen}$ and so on. One list\(^3\) for
each inflectional class, which each contains all the possible inflectional suffixes for
that category, is utilized to determine if the word has an inflectional suffix. For
instance, the list for the inflection category that $\text{stol}$ and $\text{kopp}$ belong to, thus
contains $[\text{en ar arna}]$. To find the grammatical suffix of the word, if there is one,
the ending of the word is matched to all the suffixes in each inflectional class. If
the word ends with any of the suffixes in one of the categories, it is confirmed that
all the other suffixes of that category can be attached to the root of the word,
giving rise to a correct word, i.e. a word contained in the word list. If that is the
case, then the ending of the word which matched one of the suffixes is marked as
a suffix. If, instead, not all of the suffixes give rise to a word in the word list when
attached to the root, the matching continues to the next inflectional class. The
suffix $\text{ar}$ is, for example, marked out in the word $\text{stolar}$, since it is found in the
inflectional class consisting of the suffixes $[\text{en ar arna}]$ and since attaching the
other suffixes to the root gives rise to the correct words $\text{stolen}$ and $\text{stolar}$. The
suffix $\text{ar}$ is, however, not marked out in the word $\text{promenerar}$, since the word
$\text{promener\text{en}}$ is not a correct word.

The advantage of using a word list to confirm that the part of the word preceding
the suffix is a correct word is that there is only a small risk that endings of root
morphemes are marked out as inflectional suffixes. There is, on the other hand,
the disadvantage that some inflectional suffixes are not marked out as suffixes since
not all inflections of the word is contained in the word list.

5.1.2 Determining boundaries between root morphemes

An attempt to determine whether the word is a compound or not was implemented
as the second step in the grapheme-to-phoneme module. If the word is considered
to be a compound, the boundaries between the root morphemes in this compound
are also predicted.

Rules for forming a Swedish compound word

Almost every content word in the Swedish language can be combined with other
words to form a compound. Most words can be combined through a simple con-
catenation; the words $\text{hög}$ and $\text{hus}$ can for example be concatenated into $\text{höghus}$.
However, in order to combine some words into a compound, an $s$ has to be put

\(^3\)This list was obtained from the book 'The Morphology of Present-Day Swedish'[1]
between them; the words gärd and plan are for example combined into the word gärdsplan. Some of the words that end with a or e lose this a or e when a word is concatenated to it. The words måne and sken are, for example, combined into månsken and the words lampa and skär are combined into lärskär. In addition to these rules for combining words, a few old genitive forms are also utilized, which nowadays are not used in the language except for forming compounds. When, for instance, the words kyrka and gärd are combined, the old genitive form kyrko is used, thus resulting in the word kyrkogärd.[8]

**Determining whether a part of the original word is a correct word**

All the methods mentioned above for concatenating words into a compound were considered when implementing the predicting of morpheme boundaries, except for the compound words that use old genitive forms. It was thus taken into account that when concatenating two words, an ending e or an ending a might be removed, or an s might be added.

In the implementation of the dividing of a word into root morphemes, which is described later, it is often necessary to determine whether a part of the original word is correct or not. This is determined by looking up the part of interest of the original word in the word list, and if it is found there it is considered to be correct. If, for example, the part hög in höghus is looked up in the word list, it is found, and consequently hög is considered to be a correct Swedish word.

If, however, the word is not found in the word list, it is also determined whether the word with an a or an e concatenated to it results in a word contained in the word list. Also in that case, the part of the word is considered to be a correct word. For example the grapheme combination lamp, which is a part of the word lärskär, does not exist as an independent word. It will, however, anyway be classified as a correct Swedish word in the dividing process, since adding an a to it results in the word lampa which is a Swedish word contained in the word list. To minimize the risk that incorrect words are classified as correct if concatenating an a or e by chance gives rise to a correct word, it is also confirmed that attaching other endings also yields correct words. Before confirming that the word lamp is correct, it is thus also checked that attaching the suffix or also gives rise to a correct word, in this case the Swedish word lambor.

If a part of a word is not found in the word list, and that part starts with an s, then this s might be a binding s, provided the part that is looked up is not the start of the original word. In that case, also the part of the word with the s removed is looked up in the word list. In the word gärdsplan, for example, the grapheme combination splan is not a correct Swedish word, and thus not found in the word list. However, since it starts with an s, it is also controlled whether the word following the s is a correct word, and since plan is found in the word list,
the grapheme combination slan will be classified as correct. This has, however, the disadvantage that some words inaccurately will be classified correct.

The minimum allowed length for a root morpheme

In the implementation of the dividing into root morpheme, the minimum allowed length for a root was set to three graphemes. This has the disadvantage that the words consisting of less than three graphemes are not split up correctly. The word gökur ought, for example, to be split up into gök and ur, but it will not be split up at all, since ur consists of only two graphemes. The advantage is however that a lot of incorrect dividings of words will be avoided, since there are many words that by circumstance consist of the graphemes of two-letter words without being a compound word. The word sele is for example not a compound word, nevertheless by chance consisting of the two two-letter words se and le. Even having three graphemes as the minimum allowed length can have the effect that words that are not compound words, accidentally are split up into components.

The process of dividing a compound word into root morphemes

As mentioned in the previous section, the first step in the module is finding grammatical suffixes for the word, and only thereafter the remains of the word, that is the root in which no grammatical suffixes is contained, is divided into its root morphemes. In the word bokbordet, it is, for example, first determined that et is a grammatical suffix, and thereafter the remains: bokbord is split up.

The dividing starts by determining whether a word that consists of the first three graphemes of the original is a correct root morpheme. If the first three graphemes form a correct root morpheme, then the dividing process starts recursively over again. This time, however, it is instead trying to divide the rest of the word, that is the part following the first detected root morpheme, into morphemes. This dividing can either be successful or unsuccessful. Successful means that the rest of the word can either be correctly divided into morphemes, or that the rest of the word is one correct word, given the criteria for a correct word stated above. If it is successful, then the process is completed, since a correct dividing is determined: The first morpheme consists of the first three letters of the original word, and the dividing of the rest of the word, if there is a dividing, is also known.

If either the word consisting of the first three graphemes was not a correct word, or if the dividing of the rest of the word was unsuccessful, then the process starts over. This time, however, it determines whether the first four letters is a correct first root morpheme of the original word. If it is not, the search continues with the first five graphemes, and so on. This continues until either a correct dividing is found or the length of the rest of the word goes below the minimum allowed
length of a word.

The example word barnboksaffärernas is divided into the following root morphemes: barn, bok|s and affär|ernal|s where | denotes a boundary to inflectional suffixes. The dividing is conducted in the following way:

First it is determined that the s which ends the word is an inflectional suffix and then that also ern|a is a grammatical suffix. Thereafter the actual dividing starts. The first step in the dividing process is to determine whether the first three letters of the original word is a correct word. Since the first three letters form the grapheme combination bar, which is a correct Swedish word, the dividing process starts over. This time, however, it is determined whether the rest of the original word, i.e. nboksaffärernas, either is one correct word, or a word that can be split up into several root morphemes.

The dividing process thus determines if the first three letters of the word, nb|o form a correct word. Since this is not a correct word, the process starts over by determining whether nbok is correct. Neither nbok is correct, thus nboks is the next to be checked. This continues until nboksaf is determined to be incorrect. After that point the first root morpheme is not allowed to be longer, since only two letters, är, remain in the rest of word, and the minimum allowed length for a root is two letters. It is then determined whether either nboksaff|är or nboksaffärernas is one correct word. Since neither of them is a correct word, the process terminates as unsuccessful.

The process then starts over again with the original word, but it now tests whether the first four letters form a correct word. The letters form the grapheme combination barn, which is a correct word, and therefore the dividing process starts over once again, this time however with the remaining word boksaffärernas. It is thus first tested if the first three letters of this word form a grapheme combination that is a correct word. Since the combination bok is correct, the process starts over again, but with the remains of the word, thus saffärernas.

Next, it is tested if the word saf is a correct word, and determined that is not.⁴ There is no need to test whether saff is a correct word, since the remains of the word would then only contain the two graphemes är, which is to short to be allowed. It is thus instead tested if either the entire word with inflectional suffixes, saffärernas, or the entire word without the inflectional suffixes, saffär is correct. Neither of these grapheme combinations is correct, but since the word starts with an s, this s might be a binding s. Therefore, it is also determined whether either affär or affärernas are correct words. Since at least one of them is a correct word, the dividing process terminates as successful, stating that saffärernas is one correct word.

⁴Even though the word starts with an s it is not determined whether af is correct, since the minimum allowed length for a root morpheme is set to three graphemes.
This has the effect that also the process of dividing boksaftärernas is successful, since both bok and saftärerna were classified to be correct words. Consequently, the process terminates successfully, stating that boksaffärerna consists of the root morphemes bok and affärerna, with a linking s between them.

Again, this has the effect that the process of dividing barnboksaftärernas is successful, since both barn and boksaffärernas were classified as correct. The process thus terminates successfully giving the result: barn bok|s affär|erna|s

After the dividing into morphemes is completed

If the process of dividing the original word results in two or more root morphemes, then each of these morphemes is treated as a separate word, independent of the other root morpheme, during the rest of the grapheme-to-phoneme conversion. At the end of the conversion, however, the phonetic transcriptions of the root morphemes are again concatenated into one transcribed word, thus yielding a transcription of the original word.

5.2 Determining word stress

5.2.1 Lexical stress versus phrase stress

Not all words are equally stressed in natural speech. Content words, for example nouns and verbs, are often stressed, whereas function words, for example prepositions, conjunctions and pronouns are rarely stressed.

In order to prevent that a stress is put on the most usual function words, a very short list of some common function words is utilized when synthesising the voice. If a word is contained in that list, no stress is put on it by the synthetic voice.

Where stress is put in a phrase is, however, also dependent on its semantics. Words that convey new information or are important in the sentence tend to be more stressed than other words. Therefore, also function words might in some contexts be stressed. Due to the difficulty of semantic analysis of texts, however, no attempt was made to implement this variation in stress.[6]

5.2.2 Rules for stress

The number of stressed syllables in a word

A word that is pronounced lexically can either have one or two stressed syllables. When there are two stressed syllables in a word, a primary stress is put on the first
of these, and a secondary stress is put on the second one. It is only in these primary and secondary stressed syllables that long vowels or long consonants occur.

Words that consist of only one root morpheme usually have only one stressed syllable. One example of this is the word katten, which consists of the root morpheme katt and the inflectional suffix en, and in which the syllable containing a is stressed.

Compounds, thus words consisting of more than one root morpheme, have, with a few exceptions, two stressed syllables. There are also some words, called false compounds, which have two stressed syllables even though they consist of only one root morpheme. Neither compounds with only one stressed syllable, nor false compounds were considered when implementing the rules for determining stress.[6]

**Stressable and unstressable morphemes**

In regards of stress, there are two kinds of Swedish morphemes: stressable and unstressable morphemes.

Ustressable morphemes never have a stressed syllable, independent of their context. All inflectional morphemes and some derivational morphemes are un-stressable.

Stressable morphemes are possible to stress, i.e. a stress can be put on one of the syllables of the morpheme. In some contexts these morphemes are thus stressed, and in other contexts they are not. If a word contains one or two stressable syllables, then a stress is put on all of them. If the word, instead, contains more than two stressable syllables, then a stress is put on only the first and the last of these, since a word is allowed to have a maximum of two stressed syllables.

**Stressable syllables**

In the model for word stress in Swedish described by Gösta Bruce[9], the stress in a non-compound word is always put on the last syllable of the word stem if it either ends with a consonant or with a long vowel. If the stem, instead, ends with a short vowel, the stress is positioned in one of the previous syllables, depending on if the previous syllables end with a consonant or a vowel.

However, a simpler model for stress was used, in which the first syllable of the word is the only stressable one, except if the word starts or ends with any prefixes or suffixes that effect the stress.

The prefixes that affect stress can be divided into two categories, stressable and unstressable prefixes:

- The most common unstressable prefixes are be and för. A stress is never put on a syllable in the unstressable prefix, but instead on a syllable or on
syllables in the word following the prefix. In the word betona a stress is thus put on the syllable containing o, and the in word betonbar on the syllables containing o and a.

- Most prefixes are, however, stressable, which means that they can contain a stressed syllable. The prefixes o and upp are examples of stressable ones. In the words otur and upprorisk the prefixes o and upp are therefore stressed.

Also the suffixes can be divided into two categories: Suffixes of type I and suffixes of type II. However, both these categories of suffixes are stressable.

- The suffixes of type I are stressable in the same way as stressable prefixes. Thus, they are always stressed, provided that the rest of the word does not have more than one syllable that is stressable. If the rest of the word has more than one stressable syllable, the described rules for words with three or more stressable syllables apply. Examples of suffixes of type I are the morpheme -aktig, in which the syllable containing a is the stressable one and the morpheme -het. The word läraktig has two stressable syllables and therefore both of them are stressed. In the word läraktighet, which has three stressable syllables, only the first and the last ones are stressed, i.e. the syllables containing å and e.

- A suffix of type II is not only stressable, but its presence at the end of a word has also the effect that all syllables of a preceding stressable morpheme become unstressed. Thus, if the word ends with a suffix of type II, then this suffix will have the only stressed syllable in the word, provided that the preceding part of the word does not consist of more than one stressable morpheme. The suffix -ist is of type II and therefore -ist is the only stressed syllable in a word like pianist.

If another stressable syllable is added to the word, then the preceding element contains more than one stressable morpheme. This first morpheme can then consequently also be stressed. For example, in the word superpianist the syllable containing u will be stressed, since super is a stressable prefix. In the element pianist, however, the suffix -ist will be the only stressed syllable.

For compound words, the same rules as for non-compound words apply, but the rules are instead applied to each element word of the compound individually. In the compound hundmatten, for example, the syllables containing u and a are stressed, since the syllable containing u is the only stressable in the first element, hund, and the syllable containing a is the only stressable in the last element, matten. As usual, if applying the rules individually to each of the elements of the compound word results in a total of more than two stressable syllables, then stress is put only on the first and the last of these syllables.[6]
Rules for stress that were not implemented

In general, it was not taken into account that there are some root morphemes in the Swedish language that have two stressed syllables. A few words containing these root morphemes were, however, added to the list of exceptions, for example the words äventyr and arbeta. Lists containing the most usual of the suffixes and prefixes that affect the stress of a word were used for the implementation. Far from all of the suffixes and prefixes affecting stress were, however, included in these lists. Neither was it considered that there are words to which this more simple model does not apply, for example the word valuta.

5.2.3 The implementation of detecting the stressable syllables

The process of determining stressable syllables indicates stress by putting a 1 after a vowel that is contained in a stressable syllable, and leaves all the other graphemes unchanged. The process was implemented as follows:

If the word starts with one of the unstressable prefixes, then this prefix is ignored. The process of determining stress then starts over from the beginning. This time, however, stress is determined in the part of the word that follows the unstressable prefix.

If the word, instead, starts with one of the stressable prefixes, the stressable syllable in this prefix is marked out as stressed. Also this time, the process of determining stress then starts over from the beginning, determining stressable syllable in the part of the word that follows the prefix.

If the word ends with a suffix of type I, the stressable syllable of this suffix is indicated, and thereafter stress is also marked out on the first syllable in the part preceding this suffix.

The last type of affix that can be detected in a word is a suffix of type II. If the word ends with that kind of suffix, then the stressable syllable in this suffix is marked as stressed. No other syllables are marked as stressed.

If no prefixes or suffixes affecting stress are found, then a stress is marked out on the first syllable of the word.

When the letter-to-sound rules are applied later in the grapheme-to-phoneme conversion, all vowels followed by a 1 are transcribed as stressable vowels, which is done by adding a 1 after the phonetic symbol. In the syllabification process, which thereafter takes place, this notation is utilized for determining stress.
5.3 Determining if a word has accent I or accent II

5.3.1 Swedish word accent

Swedish words have either accent I (single-tone) or accent II (double-tone). A speaker of Swedish thus uses one tone curve for pronouncing words with accent I and another tone curve for words with accent II. The realization of the tone curves for the two types of accents vary among different Swedish dialects. The form of the tone curves in the Stockholm dialect is described in section 6.3.1.

5.3.2 Rules for determining if a word has accent I or accent II

If a Swedish word has accent I or accent II can be determined through its morphologic structure, according to the following rules:

- Words with two stressed syllables have accent II. For example, the words dagbok, which is compound word and utur, which has a stressable prefix.

- Words that end with a primary stressed syllable have accent I, that is all monosyllabic words and all words with a stress on the final syllable, for example hast and kanin.

- Non-monasyllabic words that have one stressed syllable that is not the last syllable can either have accent I or accent II, though accent II is most common. Which accent it has can, to some extent, be determined through its affixes. There are six types of such affixes:

  - Blocking prefixes. A word starting with a prefix that blocks accent II has accent I, regardless of the other structure of the word. Examples of such prefixes are be- and för-.

  - Blocking suffixes. A word ending with a suffix that blocks accent II has accent I, regardless the other structure of the word. Examples of such suffixes are -is -isk -erst.

  - Inflectional suffixes which give rise to accent II. A word ending with such a suffix has accent II. The plural suffix -or and the suffix for past tense -de are examples of such suffixes, but most inflectional suffixes are of this type.

  - Derivational suffixes which give rise to accent II. A word ending with such a derivational suffix has accent II, e. g. -ska and -ling.
– Inflectional suffixes which do not affect the accent of the word. A word ending with such a suffix has thus accent I if the part of the word preceding the suffix has accent I, and accent II if the preceding part has accent II. Examples of such suffixes are the comparative suffix -\textit{re}st and the superlative suffix -\textit{erst}.

– Finally, derivational suffixes which do not affect the accent. The suffixes -\textit{est} and -\textit{iker} are examples of such suffixes.

- There are also suffixes of type II, containing two syllables, that are pronounced with a double-tone. The preceding part of the word is thus un-stressed, and the suffix is pronounced with accent II. Example of such suffixes are -\textit{essa} and -\textit{osa}.

Words that do not have any of the listed affixes, and that do not have a stressed final syllable, normally have accent II.

As always, there are a lot of exceptions to the general rules, which were not implemented. Some of the words ending with -\textit{el}, -\textit{en} or -\textit{er} have accent I and others have accent II. It is sometimes possible to predict if such a word has accent I or accent II from its morphologic structure, but not always. Neuter nouns and adjectives ending with -\textit{el} and -\textit{er} in base form have, for example, always accent I, for instance the noun \textit{tegel} and the adjective \textit{dunkel}. The non-neuter nouns that have accent II are a relatively small group, denoting kinship terms, for example the word \textit{moder}. Another example of what can be predicted is that verbs in present tense ending with -\textit{er} have always accent I.\footnote{Except for the neuter noun \textit{paper} which has accent II.}

### 5.3.3 The implementation of determining if a word has accent I or accent II

The detection of accent I and accent II words was implemented to consist of two parts: Firstly, in the part that determines inflectional suffixes, it is determined if a word ends with an inflectional suffix which gives rise to accent II. Later, it is determined if the word has a derivational suffix that gives rise to accent II and if it has an accent II blocking affix.

As described in section 5.1.1, it is determined if the word has any inflectional suffixes. However, when searching for suffixes it is also determined if this suffix is one of those that yield accent II or not. If that is the case, a 2 is added after the vowel in the suffix, to indicate that it is a suffix giving rise to accent II.

Later, after stress has been marked, the derivational suffixes giving rise to accent II are also marked. First, however, it is determined whether the word
contains any affixes that block accent II. If that is the case, it is not checked if the word has any derivational suffixes giving rise to accent II. Moreover, if any inflectional suffixes yielding accent II have previously been marked out, these markings are removed. If the word does not contain any blocking affixes, it is determined whether it has any derivational suffix that yield accent II. If that is the case, this is marked out by adding a 2 after the vowel in the suffix that has the second accent II pitch rise.\footnote{See section 6.3.1.}

It was thus not taken into account that also most non-compounds that do not have a stressed final syllable have accent II, even if they do not have any of the suffixes that give rise to accent II.

5.4 Determining if a vowel is long or short

5.4.1 Rules for whether a vowel is short or long

The main rule dictating if a vowel is to be long or not is the following: A vowel in a stressed syllable is long, provided that it is not followed by more than one consonant in the same morpheme.

The following are examples of how this rule is applied: The o is pronounced as a long vowel in all of the following words: sko, skola, skon, skons, skoldag. The o is thus stressed and either followed by one consonant or no consonants within the same morpheme. In skolnls, o has no following consonants within the same morpheme, since n and s are both inflectional suffixes and thus belonging to another morpheme. In skoldag o is followed by one consonant within the same morpheme, since skoldag is a compound word, and the d thereby belongs to the root morpheme dag.

The a is pronounced as a short vowel in the words hoppa and katt. In hoppa since it is positioned in an unstressed syllable, and in katt, since it is followed by two consonants in the same morpheme.

To this main rule there are a lot of exceptions, and most of these, but not all, are loan words. The following are, however, examples of common cases in native Swedish words in which this main rule is not valid:

In a word that ends with a single n or a single m, a preceding stressed vowel is not always long. The words rum, dröm and vän are examples of words in which the vowel preceding m and n is short.

Vowels followed by the consonant combinations rd rl or rn are usually long. This is sometimes also true for the combination rt. The words färd, pär, barn and tårta are, for example, all pronounced with a long vowel.
Some of the combinations of a consonant and one of the consonants j l n or r are also normally preceded by a long vowel, for example the words stödja, segra, tävla, särna. This is, however, not valid for all these combinations, for example not for the combinations double l, n or r.[6]

There are some exceptions to the general rule that a consonant not belonging to the same morpheme as the vowel does not affect the length of it. This exception is denoted boundary reduction. One example of such a boundary reduction is the ending t in some adjectives in neuter form. In the word högt, for example, the t influences the ö, which thereby is pronounced as a short vowel. Normally, however, it is only in monosyllabic adjectives ending with d, or t that a boundary reduction occurs. For example the adjectives söt and röd that in the neuter form becomes sött and rött, which both are pronounced with short vowels. In words like runt and vilt, that are neuter forms of adjectives ending with d, i.e. the words rund and vild, the vowel is also short. This is however not boundary reduction, but instead an effect of that the original d is deleted when the suffix denoting neuter form is added to the word. The vowel will however keep its original pronunciation, which is short.[2]

Another exception to the rule that the length of a vowel is only affected by the consonants belonging to the same morpheme is the inflection te and t of verbs, which sometimes makes the preceding vowel short. An example of this is the word köpte, in which the inflectional suffix te, affects the ö through making it short.

A vowel is also usually short when preceding a j or a x, and long when preceding a v.

5.4.2 Which of the rules for short and long vowels that were implemented

The strategy was to implement the most usual rules for long and short vowels, and to add some of the more usual exceptions to these rules to a very small transcription lexicon. The following was implemented or added to the lexicon:

- A vowel followed by any of the combinations rd, rl and rn is marked out as a long vowel.
- Some of the more common words, in which a long vowel precedes rt was added to the lexicon of exceptions.
- Vowels followed by a consonant and thereafter j, l, n or r are never marked out as short, even though some of these consonant combinations give rise to a preceding long vowel. Vowels followed by the combination of m and any of these consonants and vowels followed by double l n or r are, however, marked out as short.
• To cover for boundary reduction, all vowels preceding a double \textit{t} are marked as short, independent of whether the last \textit{t} belong to the same morpheme as the vowel or not. This has, however, the effect that vowels in words without boundary reduction inaccurately are marked as short.

• In adjectives ending with \textit{d} in the basic form and \textit{t} in neuter form no boundary marks out the \textit{d} or \textit{t} as suffixes. Thereby the \textit{t} or \textit{d} belongs to the same morpheme as the preceding vowel, and consequently the vowel in words as \textit{vild}, \textit{viit}, \textit{rund}, \textit{runt} is marked out as short.

• Vowels preceding \textit{j} or \textit{x} are marked out as short, whereas vowels preceding \textit{v} are never marked out as short.

• Some of the more common words, in which a short vowel precedes a single \textit{m} or \textit{n}, were added to the lexicon of exceptions.

5.4.3 The implementation of the rules for short or long vowel

The rules for short or long vowel were implemented as follows: It is determined whether the vowels are long or short, after it has been predicted which the stressable syllables are. It needs not to be determined if unstressed vowels are short or long, since they are always short.

For all the stressable vowels in the word, it is thus determined if the vowel is to be transcribed as a short one. If that is the case, a < is positioned after the 1 that follows the vowel, in order to mark out that the vowel is short.

If the stressed vowel precedes any of the described combinations of two consonants that normally follow a long vowel, or if the stressed vowel precedes a \textit{v}, then the vowel is not marked as a short one. Also, if the vowel is followed by only one consonant, it will not be marked as short. However, if the vowel instead is followed by a \textit{j}, an \textit{x} or two consonants it is marked out as a short vowel.

5.5 The actual letter to sound conversion

There are many rules for how different combinations of graphemes are translated into phonemes. One example of such a rule, is that \textit{k} in a front morpheme position, preceding a soft vowel it is transcribed as /\textit{c}/, and another example is that the combination \textit{rg} ending a morpheme is pronounced as /\textit{rj}/. Most of these rules have been implemented. There are, as always, many exceptions to these rules, and some of the words that do not follow the general pronunciation rules have been added to the lexicon of exceptions.
There is support for converting from a string of text to phonemes in the scripting code included in Festival, and the implementation of conversion is therefore straightforward.

All vowels followed by <1 are transcribed as short, stressed vowels, and those followed by 1, are transcribed as long, stressed vowels. The rest of the vowels are transcribed as short, unstressed vowels. The symbol | is added at the front and at the end of a word, before it transcribed, in order to mark out its start and ending. The same symbol is utilized to mark out the boundaries to suffixes and prefixes in the root morpheme.

The pre-processing of the word omkőrd, for example, transforms it into the string lo1kó1rd, in which the following rule for that transcribing the k into /ç/ applies: ( [ | k ] Vs = C ) Vs, means a soft vowel.

5.6 Syllabification

In order to later be able to model the pitch contour and the duration, the transcribed word must be divided into syllables. The following, very approximate method for dividing a word into syllables was implemented: A word is divided into one syllable for each vowel in the word. If there are at least two consonants following a short vowel, these two consonants will be associated with the same syllable as the vowel, and the rest of the consonants, if there are any, will be associated with the same syllable as the vowel following them. If a vowel is long or if there is only one consonant following it, this consonant is associated with the same syllable as the vowel. The rest of the graphemes are associated with the following syllable. The reason why this is only approximate, is that consonants, with this dividing, often will belong to the wrong syllable.

5.7 Concatenating the subwords and adjusting the accent

Up to this point, every component in a compound was implemented to be converted into phonemes separately. After the syllabification has taken place, however, the transcribed components are concatenated, yielding a transcription of the original word.

In the transcribed words, several vowels might have been marked as stressable. As described in section 5.2.2, however, a word can have a maximum of two stressed syllables. After the components of a compound have been reconcatenated it must therefore be determined on which of these a stress is to be put. Moreover, even though stressable syllables as well as suffix yielding double-tone have been marked
out, there has not yet been any co-ordination between stress and the suffixes for
double-tone. As described above, not only a double-tone suffix leads to a double-
tone word, but also two stressed syllables in a word. In addition, a double-tone
affix will not have the rise in pitch if it is not following a primary stressed syllable.
After the subwords have been concatenated into one word, the markings for stress
and accent have thus to be adjusted. Firstly, some markings for stress are removed.
The stressed syllable in a word with single-tone is marked out with 1. The primary
stressed syllable in a double-tone word is marked out with a 2, and a secondary
stressed syllable in a double-tone word is marked out with a 3. Finally, one syllable
in a suffix which leads to a double-tone in the word is marked out with a 4. This
is was implemented in the following way:

- In a word with more than two stressable syllables, the indication for stress is
  removed from all these stressed syllables, except from the last and the first
  ones in the word. This is in accordance with the rule that a stress is put
  only on the first and the last of the stressable syllables.

- In a word with two stressed syllables, the first one is primary stressed and
  the last one is secondary stressed, and the word is a double tone word. To
  indicate this, the first syllable is marked out with a 2 and the second syllable
  with a 3.

- If the word has two stressable syllables and there is also a marking for a
double-tone suffix, that double-tone suffix does not follow a primary stressed
syllable and will thus not have a rise in pitch. Therefore the marking for
double-tone suffix is removed.

However, if the word has only one stressable syllable the double-tone suf-
fix follows directly after the primary stressed syllable, and the marking for
double-tone is therefore not removed. The stressable syllable is then marked
out with a 2, to indicate that it is a primary stressed syllable in a double-tone
word, and the syllable in the double-tone suffix that has a rise in pitch is
marked out with a 4.

- If the word has only one stressable syllable and no double-tone suffix, the 1
indicating stress remains, but now it is also indicating that it is the stressed
syllable in a single-tone word.

5.8 Post-lexical processing

Pronunciation rules that reaches over word boundaries or root morpheme bound-
aries are not handled in the letter-to-sound conversion described above. The reason
for this is that each word is transcribed separately by the letter-to-sound rules, not considering the surrounding words. Such pronunciation rules are therefore, instead, handled in the post-lexical processing.

One pronunciation rule that, in my dialect, reaches over word boundaries is that r followed by t, d, s or n is pronounced as /t/, /d/, /s/ or /n/, respectively. The letter-to-sound rules, however, transcribe the r ending a word into /r/, and the t, d, s or n starting a word into /t/, /d/, /s/ and /n/. Any assimilation due to adjacent words is thus not considered. The post-lexical rules were therefore implemented to change each occurrence of the phoneme /r/ followed by the phoneme /t/, /d/ /s/ or /n/ into the phones /t/, /d/, /s/ or /n/, respectively.[6]

Post-lexical rules were also implemented for removing phone-to-phone transitions that exist in Swedish, but for which no diphones were recorded. Diphones consisting of the phone-to-phone combination of only one phone were not recorded, therefore the post-lexical rules replace adjacent occurrences of the same consonant phoneme with one occurrence. It was not taken into account that the duration and pronunciation of double occurrences might be different from the duration and pronunciation of a single phoneme. Between occurrences of the same vowel phoneme in a row, silence is instead inserted. However, since no vowel-to-vowel diphones were recorded, silence was inserted between all occurrences of two vowels in a row. Also, the diphones consisting of the phones /ɛ/, /ɛː/, /œː/ or /œ/, followed by /t/, /d/, /s/ or /n/ were not recorded, and therefore the post-lexical rules replace these vowels by their corresponding allophone that precedes /r/, /t/, /d/ and /n/.

---

[6] Of course, provided that those letters are not included in a letter combination that matches another transcribing rule.
Chapter 6

Prosody

The prosody of speech is built up of intonation, thus accent prediction and F0 modelling, of phrasing and of duration. The quality of these prosodic components in a synthetic voice are closely related to each other, and therefore there may not be any real benefit if the quality of one these components is much better then the quality of the other two.

There are two main methods that are supported by Festival for each of these components, i.e. there are two main methods for predicting phrase break, accent, F0 and duration: Either the use of a CART tree, constructed by hand or the use of a statistical model that is trained from data. In order to train the models for predicting phrase breaks, accent, F0 and duration, data labelled on prosodic features is needed. These prosodic features are, for instance, where pauses occur, where the words are stressed, how the pitch varies and how long the phonemes are.

For some of the components a lot of data is necessary to get good results. A hand made CART tree may therefore give equal or better results than a model trained on only moderate amounts of data. However, a statistical model, trained on sufficient amount of data, generally gives better results.

CART trees constructed by hand can be simple or more complicated. A simple CART tree for phrasing, for example, could assign a phrase break after each punctuation mark, and a more complicated one could also take other facts, such as position in utterance, whether the words are function or content words and so on, into account.[3]

6.1 Phrasing

In read speech there is normally a close connection between syntactic phrasing and prosodic phrasing. That is, between longer syntactic constituents, there is usually
a prosodic phrase break, which is produced through a change in pitch, duration, intensity and voice quality, and sometimes through adding a pause. In spontaneous speech, however, the relation between syntactic and prosodic phrasing is less close.

For the constructed synthetic voice, one of the CART trees included in Festival was utilized for dividing utterances into phrases. This CART tree assigns a phrase break after each punctuation mark in an utterance. However, the punctuation marks in a text do not normally separate all longer syntactic constituents that in natural speech would be separated into different prosodic phrases. Therefore the CART tree also uses the following simple model for assigning phrase breaks within text parts not separated by punctuation marks:

In Swedish it is much more common that a function word is forming a syntactic phrase with a following content word than with a preceding content word. In the phrase jag ser en katt på bordet, for example, the function word en forms a noun phrase with the following noun katt and is therefore more related with katt than with the preceding verb ser. Accordingly, the function word på forms a prepositional phrase with the following noun bordet and not with the preceding noun katt. There is thus a relatively high probability that a syntactic phrase break occurs between a content word and a following function word. Consequently, it is also more probable that the prosodic phrase breaks take place between a content word and a following function word, at least if the text is read. Therefore, in the model included in Festival, a phrase break is added between a content word and a following function word. However, in order not to make the phrases to short, a phrase break is not inserted if it is less than six words to the next punctuation mark, or if the previous punctuation mark is less than six words away. A word is considered to be a function word if it is included in the short list of function words mentioned above. If it is not included in that list, it is considered to be a content word.[3]

6.2 Pauses in speech

Speech sounds are normally produced using the air flow that arises when breathing out, therefore pauses are needed, allowing the speaker to breathe in. These pauses are normally made between two phrases, thereby not interrupting the flow of speech. However, between most phrases no pause is made, but despite this, speech appears to be divided by pauses into short phrases. It is, instead, changes in pitch, duration, intensity and voice quality that are perceived as pauses by the listener. However, to emphasize the phrase break, an actual pause could also be added, especially when reading a written text aloud. The more important the phrase breaks are, the longer are usually these pauses.

In spontaneous speech, pauses within a phrase also take place. These pauses
are not accompanied with the other prosodic features which indicate phrase break and are therefore not perceived as such by the listener. These pauses precede words before which the speaker needs time to plan how to formulate the speech, thus before important, long or difficult words or formulations. Such pauses are naturally less common when reading written text aloud.[9]

The constructed synthetic voice adds a short pause in all phrase breaks, except between phrases separated by ., ?, and : where a long pause is added. In this model, which is the default for adding pauses in Festival, all phrase breaks are consequently emphasized by a pause, and no pauses are added within phrases. This model is thus more realistic for written text that is read aloud than for spontaneous speech.

6.3 F0 modelling

6.3.1 The pitch contour of accent I and accent II

As described in section 5.3.1, Swedish words can either be accent I words or accent II words. How the pitch contours of these two accents are realized varies among different Swedish dialects. However, since the synthetic voice was constructed to have the dialect of a speaker from Stockholm, the contour of the tone curve used in this dialect was implemented. It is thus this realization of the tone curve that is described in following paragraphs.

The prosody of words pronounced within the context of a sentence is often not the same as when the word is pronounced lexically. Thus, the tone curve of a word in a phrase is not equal to the tone curve in a lexical pronunciation, but are dependent on the position of the word in the phrase and on whether the word is in focus or not.

In accent I words, whether they are in focus or not, the pitch normally starts to rise in the syllable preceding the stressed one, and falls thereafter until the beginning of the stressed syllable. If the word is in focus, the pitch starts to rise again in the stressed syllable, and falls again when reaching the syllable following the stressed one. Naturally, the rise in pitch in the syllable preceding the stressed one can not take place when this stressed syllable is the start of a phrase, since there is no preceding syllable. Occasionally, there is also no rise in the preceding syllable when the word is in focus.

In accent II words, the first pitch rise occurs later, and instead of having its pitch maximum in the syllable before the stressed one, it occurs in the stressed syllable and falls at the end of this syllable. This first pitch rise takes place in all accent II words, whereas in accent II words placed in focus, there is also a second rise in the syllable following the stressed one. However, if the word is a compound
and therefore has a primary stressed syllable and a secondary stressed syllable, this second rise will not necessary take place in the syllable following the stressed one. Instead, the second rise will occur in the secondary stressed syllable, which might be positioned in any of the syllables following the primary stressed one.[9]

6.3.2 The implementation of the pitch contour

Since no semantic analysis of the text was implemented, it is not determined whether a word is in focus or not. Instead, all words are considered to be in focus, except those normally unstressed function words included in the list mentioned in section 5.2.1. However, these function words were implemented as unstressed and have therefore neither accent I nor accent II, but were instead implemented to have a flat tone curve. Therefore, a word has either an accent I or an accent II contour, both thus with two pitch rises, or is unstressed.

To implement the pitch contour the syllables of an utterance were divided into the following seven categories:

1. The primary stressed syllable in an accent II word, i.e. the syllable in a word in which the first accent II pitch rise occurs.

2. The secondary stressed syllable if it is a word with two stressed syllables, or the syllable following the primary stressed syllable in a non-compound word. Thus, the syllable in a word in which the second accent II pitch rise occurs.

3. The primary stressed syllable in an accent I word, that either ends an utterance or is followed by an unstressed syllable.

4. The primary stressed syllable in an accent I word, that is followed by a syllable that is not unstressed.

5. An unstressed syllable following the primary stressed syllable in an accent I word.

6. An unstressed syllable that is preceding the primary stressed syllable in an accent I word.

7. Other unstressed syllables.

Moreover, three pitch levels were established:

- Neutral, which was set to 210 Hz.
- High, which was set to 220 Hz.
- Low, which was set to 200 Hz.
The tone curves for the seven syllable categories were modelled according to figure 6.1.

### 6.3.3 The pitch contour on a phrase level

When texts are read aloud there is normally a high F0 in the beginning of a phrase, which falls to a low F0 towards the end of the phrase. The initial rise and the terminal fall could either be synchronized with existing word accents in the phrase or could be separate changes in intonation.

There are also other tonal fluctuations that could take place on a phrase level, for example a continuation rise, i.e. a sudden rise from a low to a high F0 at the end of phrase.[9] These other tonal fluctuations were, however, not implemented. Neither was the pitch contour modelled to be constantly falling from the beginning of a phrase to its end.

Instead, only a rise in the phrase initial syllable and a fall in the last syllable was modelled. This rise and fall was implemented through a varying increase and decrease of the default pitch levels in the first and last syllables. The default pitch levels, i.e. those depending on to which of the seven syllable categories the syllable belongs, are in a phrase initial syllable increased by 5 to 2.5 Hz and in a syllable ending a phrase decreased by 2.5 to 5 Hz. The pitch levels are increased more in the beginning of the phrase initial syllable and less in the end of this syllable. Accordingly, the F0 is decreased less in the beginning of the last syllable and more at the end of the syllable.
6.4 Duration

6.4.1 Rules for duration

A stressed syllable has in Swedish complementary duration. This means that a consonant is short if it, within a stressed syllable, is following a vowel with long duration. A consonant or a cluster of consonants that, within the same syllable, instead follows a short vowel has a long duration. Thus, not only the consonant immediately following the stressed vowel is affected, but also subsequent consonants within the same syllable. The vowel in the word vär, for instance, is long and therefore the following consonant /g/ has a short duration. In the word vär, however, the vowel has a short duration, and the following /g/ has therefore a long duration. The length of this change in duration is dependent on if it is a consonant cluster following the vowel or if it is only one consonant. If it is a consonant cluster, then each consonant in that cluster has a smaller change in duration, than if it is only one consonant that is following the short vowel.

A change in duration is also utilized to signal phrasing. The duration tends to be longer at the end of a phrase.[9][16]

6.4.2 The implementation of duration

A list of the average duration for each phoneme was used to determine the duration of the phones.\footnote{This was not the average duration for the speaker, but the same list of duration that was used for the mebra voice [11], was utilized.} This average duration was lengthened for a consonant following a vowel with short duration. The change in duration of consonants not immediately following the vowel was, however, not accounted for. It was neither taken into account that a consonant which is included in a consonant cluster following a short vowel has a smaller change in duration than if the consonant alone follows the vowel. A change in duration on a phrase level was also implemented. The phones are pronounced with a longer duration at the end of a phrase, than in the beginning.

The change in duration was implemented through a handmade CART tree.
Part II

Evaluation
Chapter 7

Methods for evaluation of text-to-speech systems

Three main aspects of a text-to-speech system that can be evaluated will be considered here. The first aspect is the intelligibility of the synthesized voice, the second is the comprehension of what was said by the voice and the third is its naturalness.

The intelligibility of a synthesized voice is the extent to which phonetic information is correctly perceived by the listener. This can be determined by measuring the intelligibility of words or of longer speech units, for instance sentences. One example of an intelligibility test is to present a list of rhyming words to the test person and thereafter let the text-to-speech system synthesize one of these words. The test person is then supposed to determine which of the rhyming words that was said by the synthesized voice. Another example of a test is to synthesize nonsense sentences, which, however, are syntactically correct, and let the listener determine what was said. The result of intelligibility tests can be given in percent of correctly understood words and can be compared to results of the same tests made on other synthesized or real human voices.

The comprehension of a synthesized voice, on the other hand, is a measure of the amount of semantic information that is conveyed to the listener. The comprehension of a voice can be tested by letting the test person listen to a text read by a synthesized voice, and thereafter giving the listener questions on the content of the text. In contrast to intelligibility tests, a comparison with the comprehension of the same text read by a natural voice or another synthetic voice is always needed. The reason for this is that a comprehension of a text is very dependent on its level of difficulty. When synthesized voices read texts containing many names or words that are unfamiliar to the listener, for example, they are particularly difficult to understand.

When evaluating the naturalness of a voice, objective measures such as the
number of correctly understood words or correctly answered questions can not be utilized. Instead, the test person has to form a subjective opinion of the naturalness of voice, and thereafter rank this naturalness on some kind of scale.

There is another question for the evaluation of synthesized voices, which is valid for all of the three evaluation methods mentioned above: Should it be measured if a voice is intelligible, comprehensible or natural in general, or if it is intelligible, comprehensible or natural enough for a particular application? The most natural voice is not always the most suitable one in all applications. A voice with a less natural sound can for example be better when it is important to emphasize that it is computer and not a person that speaks.\cite{15}

Depending on what the aim of the evaluation is, there is also the question if a glass box or black box test is most appropriate for evaluating the text-to-speech system. In a glass box test, each component of the system is tested separately. Therefore, the glass box tests are best suited for an evaluation aiming at detecting the weaknesses of the system, in order to improve these weaker parts. In a black box test, on the other hand, the resulting output of the system is tested. Black box tests are therefore most appropriate for a general assessment of the whole system. For example to compare the general quality of two text-to-speech systems.
Chapter 8

The evaluation

Since the main objective was to determine if it is possible to construct a text-to-speech system with the methods stated in the beginning of this report, the natural choice of evaluation method is of course a method that assess the comprehensibility of the voice. However, two different tests were performed to evaluate the constructed voice. The first test was a comprehension test, but since it proved to be difficult to design a comprehension test giving reliable results with few test subjects, a second test, a hybrid between an intelligibility and a comprehension test, was also done.

Nine persons participated in the first test, and eight in the second test. The test subjects, all speaking Swedish as their native tongue, were between 19 and 56 years old, though a majority were between the middle twenties and thirty. Most of the test subjects had none or little prior experience with listening to synthesized speech.

Since the aim of the test was to assess the output of the text-to-speech system, and not to determine the performance of each component in the system, both tests were black box tests.

8.1 The first test

In the first test, four different texts were played to the test subject. The first and the last of these texts were played using my own, natural voice and the two other texts were read by the synthetic voice.

After each text had been played, the test subject was asked to answer two types of questions about it. Firstly questions on whether a word had occurred in the text, and thereafter questions on the semantic content of the text. All questions were formulated as propositions, to which the test subject was asked to state if it was a true or a false proposition, or that he or she did not know whether
it was true or false. The four test texts were chosen from a larger set of texts, previously utilized for testing the sentence-by-sentence listening time paradigm.[17] The first two texts had a very low level of difficulty, containing easy words and short sentences. The two following texts had a higher difficulty level, since they contained longer sentences and a few more difficult words. Each test subject thus listened to one easy text and one more difficult text that was read by the synthetic voice, and to one easy text and one more difficult, read by my natural voice. However, which of the two easy texts and which of the two more difficult texts that were read by the synthetic and which that were read by the natural voice, was varied among the test subjects.

Text number 1 and 2 had the lower level of difficulty, and text number 3 and 4 had the higher difficulty level.

The questions utilized for the test were also the same as in the sentence-by-sentence listening time test, though four additional questions were added, giving rise to a total of eleven questions per text.

8.1.1 Results

For each test text and for each test subject a sum was computed. For each 'true' or 'false' given by the test subject, '1' was added to the sum if this answer was correct and '-1' if it was incorrect. Since there were eleven questions, the maximum result was thus '11' and the minimum '-11'.

Two mean values were computed for each text. One was the mean of the sums for those that had heard the synthesized voice, and the other was the mean of the sums for those that had heard the natural voice. This gave the results shown in table 8.1.

<table>
<thead>
<tr>
<th></th>
<th>Mean, synthesized voice</th>
<th>Mean, natural voice</th>
</tr>
</thead>
<tbody>
<tr>
<td>Text 1</td>
<td>4.8</td>
<td>7.8</td>
</tr>
<tr>
<td>Text 2</td>
<td>8.2</td>
<td>7.3</td>
</tr>
<tr>
<td>Text 3</td>
<td>5.6</td>
<td>7.3</td>
</tr>
<tr>
<td>Text 4</td>
<td>6.3</td>
<td>7.4</td>
</tr>
<tr>
<td>Sum</td>
<td>25</td>
<td>30</td>
</tr>
</tbody>
</table>

56
8.1.2 Discussion

The sum of the mean points for the natural voice is higher than the sum of the mean points for the synthesized voice. For one of the easy texts, the results for the synthesized voice are better than for the natural one, and for the other three texts the natural voice has better results. There is also more variation between the results for the synthesized voice than for the natural one.

However, there are many limitations and difficulties in this test. It is, for example, not possible to make paired comparisons. That is, once a person has listened to a text, he or she can not be tested with the same text again. Therefore, the difference between how much a person understood of a text, read by the natural voice and how much he or she understood when the text was read by the synthesized voice can not be tested. Different persons must instead be used to evaluate the same text. This has the effect that the results of the test are interfered by differences between test persons.

In addition, the difficulty of the questions used in the test probably varies. This leads to that the value of different test scores also varies, which might have the effect that the mean scores do not give a true picture of the situation.

The differences between the results for the natural and the synthesized voice could thus be an effect of the limitations of the test. Therefore, to have been able to draw any reliable conclusions from this test, more test persons would have been needed.

8.2 A subjective assessment

Besides answering these questions on the four texts, the test subjects were also asked to subjectively assess how much of the text that they had understood.

8.2.1 Results

This subjective assessment gave the results shown in table 8.2.

8.2.2 Discussion

For all texts that were read by the synthetic voice, except text number 2, half or more than half of the test subjects chose alternative 2 or 3. Two test subjects chose alternative 2 for text number 2, and the other three test persons assessed that they understood everything or almost everything. When the texts were read by the natural voice, however, all test subjects stated that they understood everything or almost everything of the texts.
Table 8.2: The subjective assessment. '1' = 'I understood everything or almost everything', '2' = 'I understood more than half of the content of the text', '3' = 'I understood half of the content of the text'. There were also the following two options, which nobody chose: 'I understood less than half of the content of the text', 'I understood nothing or almost nothing of the content of the text'.

<table>
<thead>
<tr>
<th>Test subject:</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
<th>7</th>
<th>8</th>
<th>9</th>
<th>Mean</th>
</tr>
</thead>
<tbody>
<tr>
<td>Text 1 (natural)</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td></td>
</tr>
<tr>
<td>Text 1 (synthesized)</td>
<td>2</td>
<td>1</td>
<td>3</td>
<td>2</td>
<td>2</td>
<td>2</td>
<td>2</td>
<td>2</td>
<td>2</td>
<td></td>
</tr>
<tr>
<td>Text 2 (natural)</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td></td>
</tr>
<tr>
<td>Text 2 (synthesized)</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>2</td>
<td>2</td>
<td>2</td>
<td>2</td>
<td>2</td>
<td>1.4</td>
</tr>
<tr>
<td>Text 3 (natural)</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td></td>
</tr>
<tr>
<td>Text 3 (synthesized)</td>
<td>2</td>
<td>3</td>
<td>2</td>
<td>2</td>
<td>2</td>
<td>2</td>
<td>2</td>
<td>2</td>
<td>2</td>
<td>2.4</td>
</tr>
<tr>
<td>Text 4 (natural)</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td></td>
</tr>
<tr>
<td>Text 4 (synthesized)</td>
<td>2</td>
<td>1</td>
<td>3</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1.75</td>
</tr>
</tbody>
</table>

From these figures it can be concluded that the constructed voice is not fully comprehensible. Perhaps could even the conclusion be drawn that the constructed voice is quite far from being comprehensible, at least for texts above the lowest level of difficulty.

### 8.3 The second test

For the second test, 28 sentences were used. The sentences had a varying level of difficulty and a length that varied from three to fourteen words. Some of the sentences were made up, but most were found in different texts. The 28 sentences were divided into two groups, each group containing the same number of easy, intermediate and difficult sentences as the other group.

Each one of the eight test subjects first listened to the 14 sentences in one group, read by my natural voice. They thereafter listened to the 14 sentences in the other group, read by the synthetic voice. Which group of sentences that was read by the natural voice and which that was read by the synthetic was varied among the test subjects. Immediately after having played a sentence, a pause was made and the test subject wrote down the sentence, or the part of the sentence that he or she had perceived. When the test subject had finished writing and was ready to listen to the next sentence, he or she pressed 'enter' and the next sentence was played by the computer. The first sentences in each group were easy, and they thereafter became gradually more difficult.
8.3.1 Results

The errors were summed up as follows: For each content word that a test subject either had not written down or had perceived as another word, '1', was added to the sum. However, if the test subject had written down the correct word, but in an incorrect form, for example in a definite form instead of a definite, '0.5' was added to the error sum. If the test subject had misperceived any of the function words in such a way that it did not affect the meaning of the sentence, '0.5' was added to the error sum, if it did affect the meaning '1' was added.

The total error sum for all words and all test subjects was thereafter computed, which gave the following results:

- Sentences read by the synthetic voice had the total error sum: 235.
- Sentences read by the natural voice had the total error sum: 60.

8.3.2 Discussion

The limitations that were described in the discussion for the first test also apply to the second test. The same test person could not listen to the same sentence twice, which made paired comparisons impossible. Also, even though the sentences were chosen and grouped to give each group approximately the same level of difficulty as the other one, there probably still exists a difference in difficulty. Therefore, the difference between persons combined with the difference in difficulty between the two groups probably affected the result.

However, the total error sum for the synthesized voice is almost four times the error sum of the natural voice. This also indicates that the constructed voice is far from comprehensible.

The test was not extensive enough to determine why some words were understood by all test subjects while other words were not understood by any of them. Not surprisingly, common, short, content words in short sentences seem to have been easier to understand, than more difficult words in longer sentences.

The diphones consisting of vowel-to-vowel transitions were not recorded. Words such as *neandertalare* and *europæernas*, which in addition to being long and difficult also contain vowel-to-vowel diphones, were therefore not understood by any of the test subjects.

The word *markörer* is an example of where incorrect grapheme-to-phoneme conversion makes it impossible to understand. The word *markörer* is incorrectly split up into the subwords *mar* and *körer*, and since *kö* thereby starts a word it is incorrectly pronounced as */ç/.

There are, however, also words that neither were incorrectly transcribed, nor contained diphones that were not included in the diphone database, but yet were
incorrectly perceived. An example of this is the word pojken, which was correctly transcribed and which consists only of diphones contained in the diphone database. Two test subjects perceived pojken, but the two others instead perceived olyckan.

Thus, the shortcomings of the grapheme-to-phoneme conversion and the simplification when constructing the diphone database do not account for all misperceptions.

8.4 Comments on the constructed voice

8.4.1 Comments made by the test subjects

The test subjects were also asked to comment on the naturalness of the constructed voice. Some of the comments were:

- More difficult to understand in the beginning of a sentence
- Difficult to hear when one word ends and the next starts
- Sounds synthetic and unnatural
- Many words are pronounced incorrectly
- The stress often sounds incorrect
- Strange intonation
- Consonants starting or ending words sometimes disappear
- Background noise
- Strange pauses within some words
- Stuttering, as though the words would be split up into smaller parts
- The vowels have an incorrect duration

8.4.2 My subjective assessment

I find the quality of the constructed voice rather varying, but it always sounds very synthetic. For small, not too difficult, parts of an utterance, the voice sounds synthetic but okay. For other parts it sounds very unnatural, though it is still possible to understand. There are, however, also many utterances or parts of utterances that are completely impossible to understand.

I agree on the comments, given by the test subjects, but I have also made the following observations.
• The phone \( /\text{s}/ \) often sounds very bad. It sounds as though the synthesized \( /\text{s}/ \) would contain more high frequency sounds than a natural \( /\text{s}/ \).

• The phone \( /\text{r}/ \) often sounds bad.

• Some phones, especially plosives disappear when they end a phrase.

• A few diphones seem to be totally incorrect.

• Some vowels sound distorted.

• The explosions of the plosives sometimes sound much harder than the explosions in natural speech.
Chapter 9

Conclusion

The subjective assessment in the first test, and the results of the second test thus show that the constructed voice in many cases is not possible to understand. I would therefore not characterize the constructed voice as comprehensible.

The conclusion is thus that it is not possible to construct a comprehensible voice with the simple methods and the basic equipment that was utilized for constructing the diphone database.

If Swedish text-to-speech of a higher quality is to be constructed in Festival, a transcription lexicon is needed. The grapheme-to-phoneme rules would then only be utilized to transcribe words that are not contained in the transcription lexicon. They could also be used to divide compound words that are not in the lexicon, in order to be able to look up its subwords. Moreover, a more complete diphone database is definitely needed. The vowel-to-vowel diphones must be included, but also more diphones containing allophones of phonemes.

This is, however, probably not enough to construct a comprehensible voice, since the test subjects also made misperceptions that could not be accounted for by the incomplete database or by shortcomings in the grapheme-to-phoneme rules. Therefore, other improvements are also needed, e.g. a better recording equipment, an assessment of each diphone to decide if a re-recording of some diphones is needed and a better prosody modelling.

However, since the constructed voice is sometimes possible to comprehend, it is not unlikely that only small improvements in each module would give rise to, probably not a high quality voice, but a comprehensible voice.
Bibliography


<http://tcts.fpms.ac.be/synthesis/introtts.html>


<http://www.nada.kth.se/~viggo/stava/manual.html>

Short word list

Allophone “An audibly distinct variant of a phoneme.”

Alveolar “A phoneme that is articulated with the tip or blade of the tongue against the ridge behind the upper teeth.”

Aspirated “(Plosive) whose release is followed audibly by a short period in which the vocal cords are not vibrating.”

Content word “A word with lexical meaning, [as opposed to grammatical meaning].”

Fricative “Consonant in which the space between articulators is constricted to the point at which an air flow passes through with audible turbulence.” E.g. /s/ and /f/.

Function word “[A word] with grammatical meaning as opposed to lexical meaning: e.g. the and of are function words [...] in the top of Everest.”

Inflection “Any form or change of form which distinguishes different grammatical forms of the same lexical unit. E.g. plural books is distinguished from singular book by the inflection s.”

Inflectional class “A class of words or morphemes which have the same inflection or inflections.”

Morpheme The smallest part of a word that has either grammatical or lexical meaning.

Phone “A speech sound which is identified as the realization of a single phoneme.”

Phoneme “The smallest distinct sound unit in a given language.”

Retroflex “Articulated with the tip of the tongue or the underside of the tip against the back of the alveolar ridge.”
**Velar** “Articulated with the back of the tongue against the soft palate.”

**Voiced/voiceless** Produced with/without vibration of the vocal cords.[19]