Lexical and Acoustic Modelling of Swedish Prosody

Frid, Johan

2003

Document Version:
Publisher's PDF, also known as Version of record

TRAVAUX DE L'INSTITUT DE LINGUISTIQUE DE LUND 45

Lexical and Acoustic Modelling of Swedish Prosody

Johan Frid

LUND
UNIVERSITY

Department of Linguistics and Phonetics
Helgonabacken 12
SE-223 62 Lund

ISSN 0347-2558
ISBN 91-974116-3-9

Printed in Sweden
Studentlitteratur
Lund 2003

Contents

List of Figures

List of Tables

Acknowledgments

I wish to thank the following:

- Gösta Bruce, my supervisor. He provided excellent guidance, stimulating discussions, was always supportive, and helped me to improve this thesis in many ways.
- Merle Horne, who was always a source of knowledge, encouragement and inspiration.
- Per Lindblad, who kindly read the whole manuscript, and suggested numerous helpful improvements.
- All the labellers, recorders, processors and project leaders in the Swedia 2000 project, without whom large portions of this thesis would have been unrealizable. I especially thank the Lund-based group: Jonas Brännström, Anneli Nilsson, Susanne Schötz, My Segerup, Ida Thelander, Mechtild Tronnier and Marcus Uneson. The latter also read 4, helping to improve it.
- Paul Touati, Guus de Krom, Diana ter Brake, Gerrit Bloothooft, Mats Eeg-Olofsson, David House, Anders Eriksson, Sidney Wood, Johan Dahl, Marcus Filipsson and Birgitta Lastow, who all had a hand or two in the way things turned out.
- Arthur Holmer, who revised my English, and occasionally taught me some Basque (*Kaixo, ahuntz zaharra!*). Remaining errors are, of course, my own responsibility.
- Staff and co-workers at the Department of Linguistics and Phonetics, Lund University.
- Martin Frid, Akiko Frid, Karin Frid, Henry Frid, Richard Leufstedt, Hannah Thunberg, Lena Bengtsson, Johan Hedlund and Arno Wiering for just being wonderful people.
- Maria, my beloved wife.

Chapter 1

Introduction

This is a thesis about computational models dealing with different aspects of the automatic processing of prosody, in particular lexical stress, word accent patterns and the acoustics of intonation. Our target language is primarily Swedish, but the models and techniques used either have been or may be applied to other languages as well. The studies presented here share a common goal: to improve the modelling of prosody in a speech technology context.

Prosody and intonation are very important ingredients of human speech. In text-to-speech (TTS) applications, the utterances generated first and foremost need to be understandable. It is also desirable that they exhibit a certain degree of naturalness. Prosody modelling is, of course, integral in order to obtain this and is a highly prioritized as well as a very active sub-area in the field of speech synthesis. In speech recognition, prosody modelling is sparser, especially in commercial applications. However, most researchers agree that prosody is important (Batliner et al., 2001) and that it may provide the final boost that is necessary for the public breakthrough that speech recognition still is waiting for.

Modelling, in our view, is simplification, but we also think that one important aspect of models is that they may be designed so that they are able to perform a task, i.e., a model should not just be descriptive but also predictive. Prosody modelling, then, is largely a twofold business: modelling of linguistic structure and modelling of acoustic structure. This is a very common view that is evident in recent works in the field: Ross (1994), Dutoit (1997) and Batliner et al. (2001). The linguistic side of prosody modelling involves finding the relation between the conveyed linguistic message and the encoding of the prosodic information that is contained in this message. It is through linguistic prosody modelling that the prosody of an utterance gets a representation as a collection of abstract units that have communicative and grammatical functions. The acoustic side of prosody modelling is concerned with the acoustic manifestation of prosody in terms of F0, intensity, durations and sound spectrum. It is responsible for the generation and the analysis of F0 contours in terms of specifying the relation between the

acoustic realization of speech and the linguistic-symbolic prosody markers. This is important both in speech synthesis and automatic speech recognition.

1.1 Prosody for text-to-speech and reading

A very common strategy for the development of prosody models for TTS systems is to make them *conservative* Kochanski and Shih (2003), i.e., since the amount of prosody coded in text is limited, prosody has to be predicted from text. As erroneous prosody often is worse than "no" prosody (unvaried prosody), systems often attempt a neutral style of intonation, thereby limiting the variation in the generated speech. This is, however, no longer sufficient for commercial TTS. The importance of TTS quality is perhaps best illustrated by a note in recent paper by the AT&T group (Syrdal et al., 2000), that customers tend to rate TTS quality very highly when judging the overall quality of a voice-enabled system.

The prosodic rendering of a written piece of text depends on a number of features of the text; the lexical prosody of the words, punctuation, the mapping from raw text to words as well as semantic features of the text. The human reader performs an analysis of the above, comes up with a prosodic plan for the text, and issues articulatory commands to realize this plan as speech in conjunction with articulatory commands to realize the segmental content of the text. The task of the computer designed for text-to-speech (TTS) is somewhat similar. It must also perform an analysis of incoming text and realize it as speech sounds.

The starting point for reading, be it by man or machines, is the input of the words of the text to be read. The text then has to be processed in different ways: the word meanings have to be looked up, ambiguous words must be disambiguated, punctuation and abbreviated expressions must be dealt with. The goal of the human reader is, in most cases, to arrive at an understanding of the meaning of the text, or, when reading out loud, to convey this understanding. The goal of the computer is, typically, to render the text as speech, i.e., more similar to the task of reading out loud than reading for pleasure or for gaining knowledge and/or information. However, in order to produce a natural, human-like way of reading an input text, the computer has to understand some aspects of it. Van Herwijnen and Terken (2001) demonstrated that speakers are fairly capable of predicting what prosodic structure they would assign when reading text aloud. Human readers are often able to produce a natural phrasing and accentuation/emphasis pattern even for a text that they have never seen before as long as they have a fair amount of knowledge of the language that the text is written in.

Apart from the 'naturalness' aspect, a more fundamental aspect of prosody is that it increases the comprehensibility of a generated or spoken utterance. An obvious example is speech that is generated without pauses or any other means of grouping. Such an utterance quickly becomes hard to follow. Prosody is also sometimes

necessary for disambiguation; there are many examples of prosodic minimal pairs. It is clear that the communicative functions of prosody are more varied than the prototypical phonemic functions of segments (vowels and consonants). The most basic functions of prosody are: grouping/phrasing (boundary/coherence), prominence (foregrounding/backgrounding) and various discourse functions.

1.2 Prosody for speech recognition and listening

The interest of prosodic modelling for improving automatic speech recognition is continually growing. Prosody not only provides high-level linguistic information that might be hard to detect in individual words, but also redundant information that may be used to overcome faulty word recognition. Another growing use of prosody is the detection of disfluencies (Stolcke and Shriberg, 1996).

Prosody may be used to facilitate lexical search. It is reasonable to assume that stressed syllables are more informative for word inference than unstressed, and that knowing the stress pattern of a word can greatly reduce the number of competing word candidates. Hieronymus et al. (1992) studied the effectiveness of using acoustic stress to improve automatic speech recognition. In their system, the output of the first stage of recognition is a representation of probabilistically scored phonemes. A lexical search is then performed where the most probable word, given the phonemic input, is determined. By marking lexical stress on all content words in their lexicon – in effect, treating vowels in stressed and unstressed positions as different phonemes – they reduce the word error rate by 66% and the sentence error rate by 55% relative to a system without prosody. Wang and Seneff (2001) also incorporated a stress model in a recognizer for spontaneous speech. Although only a minor improvement was obtained, completely implausible word hypotheses were at least eliminated. The fundamental advantage of using prosody in lexical search is that it may help in narrowing the space for word hypothesization. By limiting the possible words to a small subset of the total vocabulary, recognition performance often improves.

Another important function of prosody in recognition and listening is phrasing. Shriberg et al. (2000) investigated the use of prosody for the segmentation of the input into sentence and topic units. Using speech rhythm, intonation features and pausing, a prosodic model performs as well as a word-based statistical language model. They find that pause and pitch features are especially good for segmenting news speech, whereas pause, duration and word-based cues dominate for natural conversation. In addition, Haase et al. (2001) report they could predict sentence boundaries with a precision of 93% using F0 and intensity measurements. Another study by Eisfelder and Hendrickson (1999) demonstrated the effect of prosody on listening comprehension. Students that listened to two different stories where the amount of prosodic involvement (pitch variation and emphasis) was varied showed

a greater comprehension of the stories that were read with a more expressive prosody. This shows that a more varied intonation in orally read stories has an effect on the amount of detail that can be recalled.

Other work by Wightman and Ostendorf (1994) has demonstrated the use of prosodic parameters for the automatic labelling of prosody. Prominence levels and boundaries are labelled through the use of duration, pitch and energy measures.

1.3 Outset

Much of the work in this thesis was motivated by the observation that the prosody, in particular the intonation, of Swedish has been relatively sparsely studied, at least in an academic context, by using techniques that use large pre-analysed databases in order to perform rule inference and stochastic modelling of linguistic phenomena. This is, to the best of our knowledge, true of the experimental studies presented in this thesis: the technique for inferring letter-to-sound and letter-to-prosody rules from a lexicon described in Chapter 4, the attempt at stochastic modelling of the intonation of Swedish content words in Chapter 7 and the prosody recognition experiments performed in Chapter 8.

1.4 Outline of the thesis

In this thesis, we will concentrate on the lexical aspects of linguistic modelling. We briefly treat the issue of text processing in text-to-speech, and then present a phonological analysis of lexical stress patterns of Swedish. This is followed by the development of a rule-based system for the prediction of allophonic segments and prosodic markers at the lexical level. In the acoustic modelling section, we will primarily deal with intonation. We survey previous work on Swedish and some major intonation models. Then we present a method for intonation generation that may be used in text-to-speech. Finally, we study the recognition of prosodic categories on the basis of parameterization of intonation patterns.

Here follows a brief description of each of the chapters. A more extensive summary is provided in Chapter 9.

Chapter 2 deals with the rendition of prosodic structure within the text processing component of a TTS system. Problems include the treatment of 'non-standard' and multi-compound words. Chapter 3 presents an analysis of lexical word stress in Swedish. We perform a metrical analysis of Swedish and then we reformulate this in an optimality-theoretic framework. This chapter is a slightly modified version of Frid (2001b). Chapter 4 describes the development of a machine learned rule-based system for the prediction of allophones and prosodic markers (stress and word accent) directly from the orthograpic-graphemic structure of text words. The

system is built using rule inference from a large machine-readable pronunciation dictionary and the rules are built using a technique called decision trees.

Chapter 5 is a survey of previous work on text-to-speech related intonation modelling of Swedish. Chapter 6 reviews some of the major models of intonation. Special attention is given to the paradigm of stylization. In Chapter 7, a model for the generation of intonation for Swedish words is presented. This chapter is a slightly modified version of Frid (2001a). Chapter 8 deals with recognition and characterization of word accent categories, prominence levels and dialects in Swedish. This chapter is vaguely based on Frid (2000) and Frid (2002), but it is substantially rewritten and much new material has been added. Chapter 9 summarizes the major findings in this thesis.

The title of this thesis deserves a final comment. We considered including words like "computational" or "automatic" in order to give the reader a hint that there is a slant towards speech technology in this thesis. This was ultimately rejected, since at least Chapters 3 and 6 are not concerned with computational matters. Similarly, the term "prosody" was preferred to "intonation", even though a major part of the thesis treats intonation proper rather than prosody in general. However, Chapters 3 and 4 deal with lexical stress and the latter is also concerned with vowel length.

Chapter 2

Text processing and prosodic structure

2.1 Introduction

Within text-to-speech (TTS), text analysis is a highly challenging problem. The reason is that texts, when expressed in the standard written form of a language, may offer an incomplete representation of the corresponding spoken form. Consider the following phrase in Swedish:

Karl XII:s bibel. 'The bible of Charles XII.'

This phrase offers many problems for a text analysis component. The first problem is to delimit the phrase into its subcomponents. This is fairly trivial in the writing systems of the Western languages, as words are separated by whitespace characters. This is, however not the case in for instance Chinese and Japanese. The second problem is posed by the string *XII:s*, where the roman numeral *XII* first has to be identified as such, and then converted to a definite form of a numeral, in this case *den tolfte* 'the twelfth', and then the genitive *:s* has to be appended as *s* so that the full text becomes *den tolftes* (the twelfth + GEN). Texts exhibit many units like this, which must be expanded. A third problem is that the words must be given a pronunciation. This often involves looking up the word in a computerized pronunciation dictionary, but may also have to be done by rule. A fourth problem is that prosody is only very sparsely indicated. Stress positions, prominent words and prosodic boundaries must be predicted from the text if the phrase is to have a natural spoken form.

On the whole, written materials contain less of these language signals than does spoken language. A TTS system, whose task is to render written texts as speech, has to supplement the imperfect written representation with those portions of language that are omitted. Since TTS systems (so far) have very limited capabilities of

learning, this information has to be provided through human knowledge and be expressed in a manner that the TTS system can understand. As the goal usually is to be able to read arbitrary texts, where the contents are unknown beforehand, this knowledge has to be expressed in the form of rules that can be applied generally.

The following examples from authentic Swedish news texts in the Stockholm-Umeå Corpus (SUC, Ejerhed et al., 1992) illustrate a few other cases with numbers, Roman numerals, year expressions, abbreviations and multi-compound words (translations are not given, as the purpose just is to illustrate such phenomena):

- Klockan 18 Moskvatid på fredagen – 16 svensk tid – befann sig 'Luna VIII' på drygt 45 000 km avstånd från jorden.
- Ett par emaljtallrikar 'troligen Limogesarbete, 1500-talstyp' med målade porträtt av Hugo Capet och Ludwig VIII klubbades för 2 200.
- Med tre mål inom loppet av fem minuter avgjordes cupmatchen mellan div IV-laget Nykvarn och div III-laget Västerås IK.
- Rent generellt kan man om hela den här epoken slå fast att det är den spanska barocken - litteraturen från Filip III:s och Filip IV:s Spanien, tillkommen under 1600-talets förra hälft - som dominerar det skrivna.
- Kostnaderna för den 100 %-iga subventioneringen uppskattas till 150000 kr.

Since the early days of successful TTS systems, (e.g., Allen et al., 1987) text analysis has often been performed in a preprocessing stage, where the primary task is to perform text "normalization". The goal of this normalization was basically to eliminate all non-standard (in many cases this was the same as non-alphabetic) forms of text and to find more word-like expressions for symbols such as digits, symbolic abbreviations (characters like %, $ and # (see Sproat, 2000)), and regular abbreviations, like *etc* and *Mr*. This method would produce "strings analyzable by the lexical analysis modules" (Allen et al., 1987, p. 16), consisting exclusively of words in alphabetical spelling. The preprocessing stage has then been followed by more linguistically oriented analysis methods, such as morphological and syntactic analysis. This paradigm, with separate modules, has been the dominating one within TTS development until very recently. As pointed out in Breen et al. (2002), the main reasons for using this approach have been ease of maintenance and modularity. Sproat (1998) adds that another reason might be that linguists have tended to think of phenomena like digit sequences and abbreviations as peripheral interests and that true linguistic analysis should deal with canonically spelled words.

As has been evident from the evolution of TTS systems, the "separatist" approach may be insufficient and lead to incorrect speech. The following examples,

taken from Sproat (1998), illustrate this. In English, the string *$5* would normally be expanded into *five dollars*. However, if used as a prenominal modifier, as in the phrase *$5 bill*, the expansion of *$5* would be *five dollar*. The spoken form of this expression is thus not predictable from the form itself. Another major example is the '%'-sign in Russian. Selecting the correct spoken form depends on several complex contextual factors, e.g., number, case and gender agreement. Again, this has the effect that the pronunciation is not derivable from the character alone. The core of the problem is that no access to information derived from subsequent stages of linguistic analysis is available in the text normalization approach.

In Swedish, these particular kinds of problems are not as common. Swedish orthography allows for generous compounding. Currencies are usually not, like in English, written by using transposition of the value part and the unit part of the expression. If they are – and then it is always a case of foreign currencies like pounds or dollars – they are used in ways where the spoken form is well defined and there is never any ambiguity. Using a monetary expression to denote a special kind of bill is done by preexpanding the digits and compounding: *tjugokronorssedel* 'twenty crowns bill', *tusendollarssedel* '$1000 bill'. The problem with grammatical endings is also smaller. In Swedish writing, it is allowed to directly concatenate a symbol (often a hyphen or colon) and a grammatical ending to the expression: *25%-ig, 25%-igt, 25%-iga, MFF:are, TV:s*. It is thus disambiguated directly in the text. However, it should be noted that this does not solve the problem for TTS, as allowing compounding in the orthography in this way increases the number of possible words, and makes it impossible to list all of them. The problem is not eliminated, rather it is postponed to later processing. This has the effect that in Swedish, the problems instead lie in the prosodic domain. It is difficult to know, on the basis of the word form alone, whether an expression is a single prosodic word, a compound word, or a combination of several prosodic words. In Allén (1970) all words where more than one type of grapheme – alphabetic, numeric or junctural – appear are called hybrid units. At the time of collection (1965), the ratio of these units in Swedish news texts was 0.95% of the tokens and 4.64% of the types in a million word corpus.

Monaghan (1993) points out that non-words often exhibit certain structures which makes them easily identifiable and that their intonational behaviour actually might be more regular since their semantic and pragmatic content is relatively predictable. He demonstrates this for five types of "anomalies": years, times, dates, number strings and abbreviations.

The task of the text analysis process is clearly not just a matter of converting non-alphabetic characters into alphabetic ones. Often it is also a question of deriving phonetic segments and prosody directly from the raw text input. This is for instance the case with numbers. There is little point in converting the digit string *23714* to its alphabetical equivalent *tjugotretusensjuhundrafjorton* 'twenty

three thousand seven hundred and fourteen' without predicting segmental and prosodic information at the same time. This is because the number of possible numbers is infinite – independent of the manner of representation. The list of alphabetical representations of numbers would be as infinite as the list of numerical representation. It is therefore not possible to list the pronunciations of all possible numbers. Segments and prosody must be predicted by rule for all strings consisting of digits.

In one sense, the same reasoning is valid for certain types of words. A word stem may combine with a finite set of suffixes and in compounds with a presumably infinite number of other word stems. It follows that the number of words consisting of alphabetical characters also is infinite. The suffixes affect pronunciation as well as stress and accent. Some stems are possibly non-productive, and they can be listed in a dictionary. However, we suggest that some processing of pronunciation and prosody must take place directly in the text analysis process.

Further support to this view comes from hybrid words. When analysing text, the most elegant solution is obtained when the same component that analyses '6' in the digit token *6* also is able to analyse '6' in the hybrid token *V6-motorn* 'the V6 engine' and in the possible tennis result *6-0* since they should return the same allophonic segment sequence: `s e k s`. However, they must also be analysed differently at a phrase or word level, since they should be realized with different prosodic patterns, In *6*, if the whole phrase is *6* it should get all the phrasal prosody markers. The '6' in this case gets primary stress, word accent, focus, finality and initiality. The '6' in *V6-motorn*, however, does not get any of these. Its role as a potential stress bearer is diminished since the 'V' and the 'mo'-syllable get primary and secondary stress, respectively. The problem is that hybrid words, compounds and certain derivations must be split up in order to eventually find their respective parts in a pronunciation dictionary. Still, they must also appear as a group in order for the prosody analysis to be correct.

Some of these problems are discussed in Liberman and Church (1991). They are concerned with the development of a probabilistic part-of-speech tagger. They state that it is impossible to list every possible date and year in a dictionary that produces probabilistic inputs for the tagger. Their solution is to develop a pattern matcher that recognizes non-standard word tokens and provides the probabilities instead.

An important step towards the integration of text normalization and linguistic analysis in Swedish TTS was taken in Lindström et al. (1993) and Lindström and Ljungqvist (1994). Apart from special treatment of the expansion of numbers and abbreviations, they pay special attention to the prosody of *collocations*. Collocations are certain combinations of words which often appear together, e.g. lexicalized phrases, particle verbs and other 'frozen' expressions, such as *mannen på gatan* 'the common man' and *fatta eld* 'catch fire'. They behave as one unit prosodically and

they need to be treated as such in order to increase the comprehensibility and the naturalness of synthetic speech. Since a traditional text analyser looks at one word at a time it would have problems in assigning the correct structure. Their text analysis system is therefore able to interact with the lexicon and thereby has the ability to analyse groups of words in order to form multi-word expressions. In some cases the analysis as a collocation unit may be wrong; it may just be a case of single-words units appearing together. In the phrase *Det är inte så svårt att fatta eld* 'It is not so hard to understand [how] fire [works]', *fatta eld* should be analysed as *fatta* 'understand' + *eld* 'fire'. Also consider the following triplet (from Anward and Linell, 1976), where different phrase prosody gives different meanings (the most prominent word in each phrase is preceded by an apostrophe):

- *'Gå på huset* 'walk on [the roof of] the house'
- *Gå 'på huset* 'walk into [the wall of] the house'
- *Gå på 'huset* 'visit the privy'

Only the last of these is a true collocation, where the phrase accent should be put on the last item. Therefore, their general structure may entertain several different hypotheses about the text structure and allows consequent modules to add information which eventually leads to the solution.

In Sproat (1998), the notion of text analysis is expanded to mean all computational processes involved in the conversion of "a string of characters from some character set (e.g. ascii)" (Sproat, 1998, p. 31) into a representation of the words in the text, together with their grammatical, morphological, phonological and prosodic features. The separation of the "preprocessing" and "linguistic analysis" functions is argued to be neither desirable nor necessary.

2.2 The place of text analysis in TTS

We will now describe the place of text analysis within a larger analysis frame, which, following Dutoit (1997), we will call *morpho-syntactic analysis*. In TTS systems, one of the motivations for morpho-syntactic analysis is to improve the prosodic rendering of the input text. Morpho-syntactic analysis helps to determine the organization of text, which is indispensable if natural or near-natural prosody is to be attained. Dutoit (1997) describes the process of morpho-syntactic analysis for TTS as consisting of:

- a *preprocessor*, whose task is to perform the mapping between the linear text input and the internal data structure of the modules to follow. All the words in a sentence are transformed into full text,

- a *morphological analyser*, where potential part-of-speech (POS) categories for each word are proposed.
- a *contextual analyser*, where words are considered in their contexts.
- a *syntactic-prosodic parser*, where the final structure of the text (its organization into phrases) is determined.

The preprocessor has to perform two different tasks. First, it has to recognize the units of text that have a non-linear relationship with the linguistic representation. This would include cases where multiple text units combine into one linguistic unit, e.g. numbers written with spaces separating the thousands and the hundreds and collocations (discussed above), but also where a single text unit expands to several linguistic units, e.g. abbreviations and telephone numbers written without spacing between the parts. Second, it may have to produce pronunciation data for the linguistic unit. The task of the preprocessor is further subdivided into the following:

- *Text segmentation*, where the strings of ASCII characters are transformed into lists of words.
- *Sentence end detection*, where the boundaries of sentences in the text are located.
- *Treatment of abbreviations.* This is usually done with a lexicon, where abbreviations and their corresponding words are listed. Ambiguities may arise, e.g. the single letter *f* may be used for frequency, feet, female, feminine, following [page] and franc[s]. They usually have to be solved by further syntactic, semantic and pragmatic analysis.
- *Treatment of acronyms.* This problem is similar to the abbreviations except that they are more difficult since the creation of acronyms is more productive and not as fixed as abbreviations.
- *Number processing.* Sequences of digits occur in different contexts such as dates, times, years, currencies and phone numbers, and are pronounced differently.
- *Idioms.* These are groups of words that form a single lexical unit, e.g. *by and large* and *Let's call it a day*, where the constituents cannot be analysed individually.

The method of text segmentation proposed uses a segmentation of input ASCII characters into sequences of *broad segmentation units* (BSUs)—taken from the following set:

- A sequence of alpabetic characters (upper or lower case).
- A sequence of numerical characters (digits).
- A punctuation mark or another special character (these are referred to as "junctural" graphemes in Allén (1970))
- A sequence of white space characters.

According to this analysis, a sentence as

> *The 22-year-old sailed past the milestone of his 100th goal for Liverpool.*

will be rendered as

> (The)( )(22)(-)(year)(-)(old)( )(sailed)( )(past)( )
> (the)( )(milestone)( )(of)( )(his)( )
> (100)(th)( )(goal)( )(for)( )(Liverpool)(.)

The preprocessor then applies a set of regular rules that rewrite the BSU list into a list of word-like units. Among the tasks that are performed, the BSU sequence "(22)(-)(year)(-)(old)" is converted into the string *twentytwoyearold* and the sequence "(100)(th)" is similarly converted into *hundredth*. Whitespaces are removed, as are punctuation characters. The resulting analysis is thus:

> The twentytwoyearold sailed past the milestone of his hundredth goal
> for Liverpool

In such an analysis, each word would then be passed on to the subsequent models for further processing in order to produce lexical pronunciation and prosody. But this is not the whole story. The point we would like to make here is that we think it is necessary to introduce some prosodic structure already at this level. The problem we face with a traditional text analysis system is that in the transformation process from BSUs into word-like units we might lose information that is important for the future correct phonetic and prosodic rendering of the text. In the case above, it would be more difficult to get the correct pronunciation and prosody of the word *twentytwoyearold* from the word-like unit only than from the BSU representation, since we have lost the information of the internal structure of the word. The only foolproof way to get a correct analysis of a word is if there is an exact match between the word and an item in the lexicon. We cannot trust the lexicon to contain this word, as it would then have to contain an enormous (infinite) number of constructions like this (e.g. *twentythreeyearold*, *twentyfouryearold* etc) so we have to rely on rules. We might use morphological decomposition methods to analyse

the structure of the word but then we might as well use the original, unconverted, components. Using the BSUs, we are able to generate prosody and pronunciation for each part individually. We then use rules for the derivation of word-internal coarticulation phenomena and prosody in complex words. If we instead try to guess these properties from the alphabetic representation we might end up – given the complexities of the letter-to-sound relationship in English – with something like `t w e n t i t w oi r ou l d`.[1]

For prosody, the situation is similar. Rendering the correct prosody in a multi-compound word is very difficult and often simplistically modelled in text-to-speech (as shown below in Section 2.5). Clues about the internal structure definitely help. A system that predicts the prosodic structure of words from alphabetic input, without the use of a dictionary, is presented in Chapter 4.

2.3 BSU analysis of Swedish texts

When the text analysis engine of a TTS system encounters character strings like *22-year-old* or *100th*, the BSU analysis would then split up the strings according to the properties of the characters. We examined the occurrence of non-standard words in a text corpus consisting of a union of the SUC corpus and another corpus, from the Swedish daily newpaper DN. This combined corpus, which we will call 'sucdn' consisted of 2291666 word tokens. We performed a BSU analysis of each word token in the corpus. Groups of alphabetic characters are labelled 'a', groups of numerical characters are labelled 'n', whereas junctural characters are shown as is. Since the analysis is performed on word-level, white space characters are not included in the analysis. The majority of cases (91.7%) are alphabetic. Another 6.5% consist of single structural characters (mostly periods and commas, SUC seems to sometimes separate the period from the rest of the sentence) and a third relatively large group are the all-numerics (0.8%). The rest of the cases (0.6%) are thus structurally complex, i.e., consisting of more than one subsequence of the types described above. Table 2.1 shows a list of examples of different BSU sequences. We have found 201 different types, 57 of which occur at least 10 times or more, and 134 of which occur more than one time. The most common type is a - a, where an alphabetic sequence is followed by a dash and then another alphabetic sequence. This is a common way of writing compounds. Other common types are n - a (*6-åringen* 'the six-year-old'), a - (*in-* as in *in- och utgångar* 'entrances and exits') and a _ a (alternative ways of writing compounds and abbreviations).

[1] This is actually the output, in MRPA transcription, of the otherwise quite robust Festival (British English) rules (Taylor et al., 1998), but other systems fail on this one as well. The literal string *22-year-old* gives the correct pronunciation, however.

Table 2.1: Illustration of the BSU analysis. Only a selection of different types is shown.

Freq. of occurrence	Type	Examples
809	a : a	ABB:s, AIK:are, UNIX:s, XII:s, S:t
730	n - n	1-0, 08-701, 1-2, 1993-1997
647	n . n	0.1, 00.17, 13.90, 15.7, 150.000, 1.0
644	n , n	0,000001; 1,5; 400,0000
515	n : n	01:15, 1947:576, 1:50000, 9:1
437	a . a	B.F, bl.a, m.fl, m.m, s.k, t.ex
306	$ a	$Ã„, $Â…
279	n / n	1/2, 1990/91, 6/92, 28/2
196	a ’ a	McDonald’s, sta’n, l’ancien, d’Orsay
172	a - a - a	S-E-banken, gröna-vågen-romantiker, fransk-spansk-italiensk
157	a _ a _ a	d_v_s, t_o_m, p_g_a
156	n : a	1:a, 4:an, 3:orna
142	a n	m3, X2, C60, V6, R4, P3
134	n . n . n	3.2.1, 2.1.4
127	n a	2a, 1992a, 20V, 525i, 9årige
100	a / a	m/s, km/tim, Dahlgren/Pramling, och/eller
97	n -	1600-, 40-
66	- n	-92
54	a - a - a - a	text-till-tal-omvandlare
49	n : n - n	1983:81-85

2.4 A taxonomy of non-standard words

The conversion of BSUs into words and phonetic information differs depending on the type of text unit. Some text units contain alphabetic graphemes, some contain numbers, others junctural graphemes. There may even be units consisting of combinations of several different types of graphemes. Before we know what type of unit we are dealing with, it is difficult to apply the correct conversion method. In Sproat et al. (2001) this problem is analysed. They develop a taxonomy of non-standard words. The taxonomy is based on several different types of text including news texts, internet newsgroup texts and advertisements. The categories were chosen to reflect anticipated differences in algorithms for transforming the text units into words. It was thus not developed specifically with prosody in mind. The taxonomy is shown in Table 2.2.

Note that the actual categorization is more dependent on how words should be

Table 2.2: A taxonomy of non-standard words. Adopted from Sproat et al. (2001)

Alphabetic	EXPN	abbreviation, expand to full word or word sequence
	LSEQ	letter sequence (say each letter)
	ASWD	read as word, both standard words and acronyms that are said as words, e.g *NATO*, *WYSIWYG* and *CAD*.
	MSPL	misspelling
Numbers	NUM	cardinal number
	NORD	ordinal number
	NTEL	(part of) telephone number
	NDIG	as digits
	NIDE	identifier (serienummer)
	NADDR	part of street adress
	NZIP	zip code
	NTIME	time
	NDATE	date
	NYER	year
	MONEY	money
	BMONEY	big amount of money
	PRCT	percentage
Miscellaneous	SPLT	mixed or split ('hybrid' units in Allén, 1970)
	SLNT	not spoken, word boundary
	PUNC	not spoken, phrase boundary
	FNSP	funny spelling
	URL	url or other internet related
	NONE	ignored

pronounced than how they are rendered in the text, e.g., there is no special class for roman numerals despite their rendering as letters instead of numbers, since they are spoken like numbers, and might be spoken like a cardinal or an ordinal depending on context. Likewise, the string "2002" could be labelled either as NUM or as NYER, depending on how it should be read.

2.4.1 Examples

We will now exemplify how prosodic patterns may be assigned for some of the numerical NSW categories. For the other categories, other algorithms for prosody assignment are applied. The prosody of acronyms, for example, has been investigated by Bruce (1993) and Lindberg et al. (1997).

Cardinal numbers (NUM)

Cardinal numbers consist of a sequence of digits. These are first converted into words from the following groups:

Group	Members
A	the words for numbers from 1 to 9: ett, två, tre, fyra, fem, sex, sju, åtta, nio
B	the words for numbers from 10 to 19: tio, elva, tolv, tretton, fjorton, femton, sexton, sjutton, arton, nitton
C	the words for the tens from 20 to 90: tjugo, trettio, fyrtio, femtio, sextio, sjuttio, åttio, nittio
D	the union of groups A, B and C
E	the words for 100, 1000, 1000000 etc: hundra, tusen, miljon etc

The digits are parsed from right to left. A single digit gets the corresponding label from group A. For all other lengths of digit strings, the two leftmost digits get one label from group B if its numerical value is less than 20, or one label from group C and one from group A. If the last digit is 0, no label from group A is assigned. Correspondingly, if the second last digit is 0, no label from group C is assigned[2]. The third digit from the right is assigned the corresponding number from group A together with the word 'hundra' from group E. The fourth to the sixth digit from the right are parsed in the same way as the first to the third, and adds the word 'tusen' from group E. The process is then repeated for all groups of three digits, adding different words from group E for each group. Note that the *ett* 'one' before 'hundra' and 'tusen' is optional.

Digit sequence	Parse	From groups
1	(ett)	A
10	(tio)	B
20	(tjugo)	C
21	(tjugo)(ett)	C + A
123	(ett)(hundra)(tjugo)(tre)	A + E + C + A
1212	(ett)(tusen)(två)(hundra)(tolv)	A + E + A + E + B
65536	(sextio)(fem)(tusen)(fem)(hundra) (trettio)(sex)	C + A + E + A + E + C + A

[2]Note that cases where the digit string starts with a 0, like in the area code *040*, or the movie-agent designation *007*, are treated by the type NDIG of the taxonomy

Stress is then assigned to the first syllable of the rightmost label that is a member of group D. These labels are in boldface below.

55 (femtio)(**fem**)
250 (två)(hundra)(**femtio**)
312 (tre)(hundra)(**tolv**)
500 (**fem**)(hundra)

If the *ett* is left out before 'hundra' and 'tusen' they are stressed if they appear as a single word, otherwise they are usually unstressed.

Note that there are variations regarding for instance speech rate and dialects. In a speech synthesis system it is, however, possible to specify how much of this variation that the system actually should include.

The cardinal numbers from 0 to 99 have one stressed syllable. The cardinal numbers from 100 to 199 also have one stressed syllable, unless the optional *ett* 'one' is used. Then the prosody depends on speech tempo. The same applies to the numbers from 200 to 999.

133	(without *ett*)	(hundra)(trettio)(**tre**)
133	(with *ett*, slow)	(**ett**)(hundra)(trettio)(**tre**)
133	(with *ett*, fast)	(ett)(hundra)(trettio)(**tre**)

Higher numbers often have several stressed syllables. Actally, each group of hundreds, thousands, millions etc may be assigned a stressed syllable. The phrases can be seen as consisting of several prosodic words.

33333333 (trettio)(**tre**)(miljoner)
(tre)(hundra)(trettio)(**tre**)(tusen)
(tre)(hundra)(trettio)(**tre**)

Years (NYER)

Prosody is assigned as for the '200-999' group above. The 'hundred' loop is continued. Again, speech tempo decides whether there are one or two prosodic words.

1345 (**tretton**)(hundra)(fyrtio)(**fem**)
2003 (slow) (**tjugo**)(hundra)(**tre**)
2003 (fast) (tjugo)(hundra)(**tre**)

2.4.2 Discussion

The idea of the BSU analysis is to determine what type of non-standard word a given token is. Depending on the type, different algorithms for the derivation of pronunciation and prosody may be applied.

There are different types of problems associated with each group. A general problem is that the mapping between BSUs and NSWs is highly ambiguous. Consider, e.g. the 'n . n' group. The instances of this group might be decimal numbers, time denunciation, version numbers, pricing information, big numbers (where the point is a delimiter of groups of thousands) and dates. Yarowsky (1996) has proposed a solution to the disambiguation problem. He used a large text database and located all occurrences of a homographic token. Each occurrence was then labelled into classes and context features were extracted. A classification tree (See Chapter 4) may then by built based on these features.

Another problem is related to the grouping of the elements in a BSU. In a category such as 'a - a - a' the instances might have several different prosodic structures. This is further examined in Section 2.5.

It is also possible to refine the taxonomy, e.g. with additional types for mathematical expressions, scales and ratios.

2.5 Prosody in multi-compound words

We shall here present a small study on the strategies adopted for assigning prosody in multi-compound words in some existing commercial Swedish TTS systems. There are basically two types of prosodic patterns above word-level in Swedish: the "lexicalized phrase"-pattern (LP-pattern, see Anward and Linell, 1976; Rischel, 1983, on unit accentuation in Danish) and the compound pattern (see Zonneveld and Bruce, 1999; Bruce, 1998). The LP pattern consists of one main stress only, always on the last element, as in *högsta 'domstolen* 'the supreme court' and *hålla 'tal* 'make a speech'. Exceptions occur, e.g. when the last element is a reflexive pronoun or when it modifies the meaning of the earlier part, e.g. indicating that something is going to be performed to a lesser degree, e.g. *göra 'bort sig* 'make a fool of oneself', *deka 'ner sig* '(coll.) go to the dogs' and *'sova en smula* 'sleep a little'. Potential stress-bearers in earlier elements of the unit are typically unstressed. There may be some variation, however, so that a secondary stress can occur on the first element of the unit.

The compound pattern consists of a main stress and a secondary stress. The main stress always comes first. Usually, main stress is found on the first stressable element and secondary stress on the last stressable element. Potential stress-bearers in the middle are usually unstressed, but in some instances it is possible to have more than one secondary stress. The compound pattern may be superordinate,

meaning that it is possible to have the LP pattern inside a compound pattern. The LP pattern may, however, also occur independently[3]. The LP pattern may be seen as right-headed (stress comes last), and may be used either on its own, or as the first part of a compound, whereas the compound pattern is left-headed.

2.5.1 Material and method

We extracted a list of ten graphemically complex words from the SUC and DN corpora. They have various internal prosodic grouping structures. The phrases are listed in Table 2.3 with their respective prosodic structures. Prosodic groupings are shown using parentheses, X signals stressed units, and . signals unstressed units. We then tested them on two different existing demonstrational versions of commercial Swedish TTS-systems: L & H (SpeechSoft) and Infovox. The idea was not to evaluate the systems, but rather to examine them. We will therefore use the anonymous labels 'System 1' and 'System 2' to denote them. After synthesising the utterances, each one was prosodically analysed.

2.5.2 Results

The results of the analysis are presented in Table 2.4. The strategy adopted by System 1 is to use the compound pattern throughout, with main stress on the first element, and secondary stress on the last one, like the phrases *C-dur-kvartett* and *a-kasse-ersättningen* (for translations, see Table 2.3). System 2 is less general. It mainly uses the LP pattern, and realizes the last item as if it were spoken in isolation, i.e. giving it compound accent if it is a lexical compound, otherwise just one stress on the lexically stressed syllable. System 2 thus, effectively, forms two prosodic words. The first prosodic word is formed by all but the last item, which then forms the second prosodic word. The strategy in both systems is to have a default pattern that is applied indifferently for this type of word.

2.5.3 Discussion

When determining the prosodic structure of a multi-compound word we must decide whether is should have compound or LP prosody, or a combination, where the whole expression has compound prosody, but where the first element is an LP. In the latter case, we must also determine the span of this LP so that main stress is correctly assigned. One step towards a solution of this problem is to analyse the internal structure of the multi-part words in order to discover if they contain a lexicaliszed phrase. Lindberg (2000) has developed a lexicon-based method for

[3]In fact, phrase prosody could very well be viewed as a succession of LP patterns, where there sometimes is a compound part on the right-hand side containing a secondary stress. Under this analysis, the compound level would be subordinate to the phrase level.

Table 2.3: Phrases in the multi-compound word test and their prosodic structure.

1. *C-dur-kvartett* 'quartet in C major'				
C-	dur-	kvartett		
((X	.)	x)		
2. *dygnet-runt-service* 'round the clock service'				
dygnet-	runt-	service		
((.	X)	x)		
3. *Europa-topp-listan* 'European top hit list'				
Europa-	topp-	listan		(avoidance of
(X	(.	x))		stress clash)
4. *a-kasse-ersättningen* 'unemployment benefit'				
a-	kasse-	ersättningen		
((X	.)	x)		
5. *berg-och-dalbana* 'roller-coaster'				
berg-	och-	dal	bana	
((.	.	X)	x)	(misspelling)
6. *brittisk-norsk-svensk* 'Brittish-Norwegian-Swedish'				
brittisk-	norsk-	svensk		
(X	.	x)		
7. *drop-in-mottagning* 'drop-in reception'				
drop-	in-	mottagning		
((.	X)	x)		
8. *dörr-till-dörr-transport* 'door to door shipment'				
dörr-	till-	dörr-	transport	
((.	.	X)	x)	
9. *fonem-grafem-översättningen* 'phoneme-grapheme translation'				
fonem-	grafem-	översättningen		
((.	X)	x)		
10. *icke-Nato-anslutna* 'not affiliated to Nato'				
icke-	Nato-	anslutna		
(X	(.	x)) *or*		
(X)	(X	x)		
11. *text-till-tal-omvandlare* 'text to speech converter'				
text-	till-	tal-	omvandlare	
((.	.	X)	x)	
12. *året-runt-avlönade* 'salaried all year round'				
året-	runt-	avlönade		
((.	X)	x)		

Table 2.4: Prosodic analyses

	System 1	System 2
1.	((X .) x)	(. X) (X)
2.	((X .) x)	(. X) (X)
3.	(X .) x)	(X) (X x)
4.	((X .) x)	(. X) (X x)
5.	((X .) x)	(. X) (X)
6.	(X . x)	(. X) (X)
7.	((X .) x)	(. X) (X x)
8.	((X .) x)	(. X) (X)
9.	((X .) x)	(. X) (X x)
10a.	(X (. x))	(. (X x))
10b.	(X) (X x)	(X) (X x)
11.	((X .) x)	(x X) (X x)
12.	((X x) x)	(x X) (X x)

detection of lexicalized phrases. Jande (2001) has examined the stress properties of these phrases. However, the problem still remains to be solved satisfactorily. Our own solution, which does not use a dictionary, is presented in Capter 4.

2.5.4 Summary

In this chapter, we have discussed a few problems related to the generation of correct prosodic structure in text-to-speech. One problem is that the text processing task is often modelled as a text *normalization* task, where anomalous text words are alphabetized. However, this often only results in a conversion from one uncountable series to another, as is the case with numbers.

We suggested a strategy that uses an analysis of anomalous words into *Broad Segmentation Units*. These units may then be mapped to certain types of *Non-Standard Words*. For each NSW there then is an algorithm that produces the desired allophones and prosodic structures.

We also performed a small study of some existing text-to-speech conversion systems for Swedish, observing their behaviour on the task of generating word prosody for multi-compound words. We illustrated that there still might be room for future improvements but also that the task is a highly complex one.

Chapter 3

Swedish word stress in metrical phonology and optimality theory

3.1 Introduction

The purpose of this chapter is to give an introduction to how lexical word stress in Swedish can be analysed with modern phonological theories such as metrical phonology (Liberman, 1975) and optimality theory (Prince and Smolensky, 1993). Central concepts and structures within the phonological theories are introduced and discussed, and examples of how the word stress pattern of Swedish can be treated within optimality theory (henceforth: *OT*) are given. We will deal with monomorphemic words, compound words and affixes.

3.2 Metrical phonology

Within OT, word stress has mainly been analysed using concepts borrowed from metrical phonology, e.g. feet and syllable weight. We will therefore first give a short introduction to this theory.

Metrical phonology is a theory about rhythm and stress in languages, and part of its roots comes from the metrical descriptions of ancient Greek and Latin poetry. However, the origin of modern metrical phonology is Liberman (1975) and was further developed by e.g. Liberman and Prince (1977). One of the distinguishing characteristics of metrical phonology is that it not only shows the relationship between different prominence levels, but also the grouping pattern, i.e. the forming of prosodic groups triggered by stress.

3.2.1 Prosodic hierarchies

In order to describe lexical word stress a prosodic hierarchy is often used. The basic form of this has three levels: syllable, foot and word. The syllables are the smallest units and carry stress. On the foot level syllables are grouped together, and one syllable within every group is identified as the head of the foot. On the word level one foot is identified as the head foot, also becoming the head of the whole word. Nespor and Vogel (1986) and Hayes (1995) put another level above the word (but below the phrase): the clitic group. We will use this level to treat the word stress pattern of compound words (see Section 3.4.6). Often another level below the syllable is used: the mora level. A mora is an abstract length unit, and is used to show the weight of a syllable. Following common procedure, we will use Greek letters to symbolise levels, e.g., small *sigma* (σ) for syllable, and small *mu* (μ) for mora.

3.2.2 Metrical grids

The rhythmical structure of linguistic units is usually illustrated in metrical grids. This is done by assigning a symbol to each unit on a given level. A strong unit receives the symbol 'x' and a weak unit the symbol '.'. In order to show the grouping structure parentheses are put around the units that are grouped together, see Table 3.1. Putting each level in the prosodic hierarchy on a line of its own shows the hierarchical structure.

Table 3.1: Metrical grid showing the prosodic hierarchy of the word *reducera*.

Level										
Word	(					x		)		
Foot	(	x	.	)	(	x	.	)		
Syllable		σ	σ			σ	σ			
		re	dɵ			se:	ra		*reducera* 'reduce'	

In Table 3.1 it can be seen that the word's primary stress is on the penultimate syllable, while there is a strong, rhythmically induced, unstressed syllable in initial position. The grouping relations are seen as well. An important principle within metrical phonology is that a strong unit on one level must be supported by a strong unit in the same column on the level below. This is what Hayes (1995) calls *The Continuous Column Constraint* (CCC).

3.2.3 Parameters

It is common to formalise the description of the metrical structure of a language by using the following five parameters (based on Hayes (1995) and Kager (1995)):

1. Boundedness: whether the language has feet with more than two syllables or not.
2. Quantity sensitivity: whether the language distinguishes between different syllable weights or not.
3. Foot headedness: where the head of the foot is.
4. Word headedness: where the head of the word is.
5. Directionality: the direction of foot formation (forwards or backwards).

Parametric analysis of Swedish

The description of stress in Swedish non-compound words in Bruce (1998); Zonneveld and Bruce (1999) yields the following parameter values: Swedish has bounded feet (the fundamental pattern is bisyllabic), is quantity sensitive (interacts with the weight of the rhyme of a syllable), the foot head is left-bounded (trochaic), the word-head is right-bounded, and the foot formation starts at the right edge of the word.

In compound words each morpheme is analysed first, then the position of the primary stress is determined.

3.2.4 Universal foot inventory

A later stage in the development of metrical theory abandons the parametrical description since they can combine to create stress systems that are either rare or unattested (Hayes, 1987). Instead, a set of foot structures is suggested, which function as theoretical primitives. These are given in Table 3.2.

Table 3.2: Foot structures suggested by Hayes (1987).

Syllabic trochee:	(	x	.	)				
		σ	σ					
Moraic trochee:	(	x	.	)	or	(	x	)
		σ_μ	σ_μ				$\sigma_{\mu\mu}$	
Iamb:	(	.	x	)	or	(	x	)
		σ_μ	$\sigma_{\mu\mu}$				$\sigma_{\mu\mu}$	

Feet used by Swedish

According to Riad (1992) the dominating foot in Swedish is the moraic trochee, which is realized either with a long vowel consisting of two morae or a short vowel + consonant, which realize one mora each. However, see Section 3.4.5.

3.2.5 Extrametrical syllables

Some syllables can not be grouped together with any other syllable and can also not form a group of their own since they only have a weak unit (less than two morae). When a syllable is weak (less than two morae), but is unable to form a foot with another syllable, the weak syllable is called *extrametric*, and is left unparsed by the foot-forming procedure. This occurs for instance, in words with two syllables and final stress, like *banan* 'banana' (if we do not accept iambic feet in Swedish). Extrametric syllables are shown within angled brackets, as in the following example:

<σ>	(x)
ba	nɑ:n

3.3 Optimality theory

Here we give a brief presentation of the structure and analysis method of OT. Central concepts such as tableaux, constraints and rankings are introduced.

3.3.1 Introduction

OT (Prince and Smolensky, 1993) is a development of generative grammar and shares with it a focus on formal descriptions and the search for universal features among the world's languages. The central idea within OT is that surface forms in languages are the result of a tug-of-war between competing grammatical principles, called *constraints*. In this way OT differs from traditional grammar, which uses rewriting or transformational rules. In traditional grammar, one form is derived from another with rules. In OT, representations are eliminated when they violate a constraint until one candidate remains, the winning or *optimal* candidate. OT thus concentrates on the interaction between grammatical principles. OT should be seen as a general theory of grammar. It has mostly been used for phonology, but the number of studies within syntax and morphology is increasing. OT, like generative grammar, claims to be a theory about the human language capacity.

Structures, concepts and analysis method

An optimality theoretic description of a linguistic phenomenon consists of an input form, a grammar (sometimes called GEN) that generates all possible output candidates from the input, and a set of constraints that decide the outcome of the grammar. The constraints are ranked, i.e., they are applied in a specified sequential order. The constraints are also universal, i.e. they are valid for all human languages. Structural differences between languages depend on different rankings from one

language to another. Since the constraints eliminate candidates as they are applied, the final remaining candidate is the winning or optimal candidate. In order to 'win' – to become the optimal candidate – a candidate does thus not need to satisfy all constraints in order to be grammatical, it suffices that it is better than all the competing candidates (for the same underlying input form). This is perhaps the greatest difference compared to traditional grammar. The mechanism that evaluates the grammar is sometimes called EVAL, or H-EVAL, where the H stands for 'harmonic', which in this case means that the candidate that is most harmonic in relation to the constraints is preferred.

3.3.2 Tableaux

Optimality theoretic analyses are often represented in tableaux. These show the input form, the constraint ranking, selected candidates and their violations, and the winning candidate. Violations are marked with an asterisk: '*' (many violations of the same constraint cause more '*'). When a constraint violation means that a candidate becomes non-optimal, i.e., that there are other remaining candidates not violating this constraint (or violating it to a lesser degree), this is marked with '!', and the candidate's fields for the lower ranked constraints are shaded. A winning candidate is shown with a pointing hand. See Figure 3.1.

/form/	CONSTRAINT 1	CONSTRAINT 2
candidate 1	*!	
☞ candidate 2		*

Figure 3.1: Illustration of OT tableaux. Each candidate is presented on a separate line, and the constraints are shown at the top of each column, with the highest ranked constraint to the left. Violations are indicated with an asterisk, fatal violations with exclamation mark, and the winning candidate with a pointing hand.

3.4 Swedish word stress in OT

Before moving into how stress is treated in OT, we shall give a brief summary of the rule system for lexical stress in Swedish. We will not deal with phrasal stress here, but we will include aspects of compound words and derivatives (root + affix).

3.4.1 The placement of stress in Swedish words

Stress is a fundamental rhythmical feature in Swedish, and it is perceptually important that stress comes at the correct position in words. Stress is a feature of the syllable, while accentuation and focussing are features of the foot and word respectively. The position of stress in Swedish words is not fixed, it can occur in different positions. This means that stress can be distinctive, i.e., two words can differ only in their stress pattern, as in *'fasan* 'the horror' and *fa'san* 'pheasant'. This often also causes a change in vowel quality. Inflectional endings making up whole syllables do never take stress. There is also a connection between stress and syllable weight: a stressed syllable is always heavy. Bruce (1993) summarises the most important rules for Swedish stress placement. A fundamental difference is made between monomorphemic and compound words (true compounds and derivatives).

Monomorphemic words

Monomorphemic words consist of one root morpheme. The following rules apply:

- if the final syllable is closed (or otherwise heavy), stress is placed on this syllable. Exception: if the final syllable is *-el*, *-en* or *-er*. These syllables often contain a /ə/ (schwa) vowel, which never is stressed.
- if the final syllable is open, and the penult is closed, stress is placed on the penult.
- if both the final and the penult are open, stress is placed on the antepenult.

As noted by Bruce, it is easy to find counter-examples. We will adopt this analysis with a change regarding trisyllabic words; a closed final syllable does not always receive stress in this case. The following principles will be used:

- polysyllabic words have penultimate stress.
- superheavy[1] final syllables have final stress.
- trisyllabic words with an open penult and closed final syllable get antepenultimate stress.
- exceptions have prespecified foot patterns in their input forms.

The following examples illustrate the principles:

[1] Superheavy syllables have three timing positions or morae in the rhyme. According to Kiparsky (2003) they occur in all varieties of Swedish.

- Polysyllabic words have penultimate stress:

a.'mø:.ba	*amöba*	'amoeba'
jɛ.'stal.ta	*gestalta*	'to shape'
'ɑ:.nɪs	*anis*	'aniseed'

- Superheavy final syllables give the word final stress:

ba.'nɑ:n	*banan*	'banana'
ka.ta.'stro:f	*katastrof*	'disaster'

- Trisyllabic words with open penult and closed final get antepenultimate stress. However, closed penult results in penultimate stress:

'mɑ:.ra.tɔn	*maraton*	'marathon'
re.'ak.tɔr	*reaktor*	'reactor'

Derivatives

Derivatives consist of a root morpheme plus affixes (prefixes and suffixes). Affixation can affect the stress pattern in different ways. The following situations occur (affixes not translated)

- affix does not affect stress pattern: *be-*, *ent-*, *för-*, *-ande*, *-else*.
- affix attracts stress and deprives the root of stress: *-ant*, *-graf*, *-ör*.
- affix attracts stress but does not deprive the root of stress; word behaves like compound (see the next Section): *hyper-*, *o-*, *-artad*, *-bar*, *-het*.

Compounds

Compound words usually have two stresses, one on the first stressable element, and one on the final stressable element. The first one of these gets primary stress, while the second gets secondary stress. Stressable elements are the root, and the affixes that carry stress. The characteristic thing about compounds is that they consist of (at least) two morphemes, each with stress. However, as compounds only have one primary stress, the 'surplus' of stresses must be solved.

3.4.2 Candidate generation

For every input form fully metrified candidates are constructed. This includes grouping syllables into feet (foot formation), assigning the head of each foot, and assigning a head foot of the word. The number of formal possibilities becomes uneconomical, since every combination of grouping, foot headedness and word headedness must be generated. Therefore, it is common practice not to show all

the candidates in tableaux, only those that best illustrate the features of the grammar or the forms attested elsewhere in the language.

What is the correct input form?

For Swedish, there is a problem in using the quantity distinction in the input forms, since the analysis becomes circular (quantity is used to predict stress, which predicts quantity differences). The input form should therefore not contain any quantity information, and hence no vowel quality information since this usually is derived from the quantity.

3.4.3 Mora counting

A fundamental unit in the following analysis is the mora. A mora is an abstract length unit and it is on the level below the syllable in the prosodic hierarchy. Syllables are usually monomoraic, but syllables that are foot heads are (at least) bimoraic. Vowels in the input form primarily count as one mora, but they can be analysed as two morae in a syllable that is a foot head. In the syllable with primary stress, and in the syllables following that one, coda consonants are also moraic. A bimoraic vowel is normally realized as a 'long' vowel, and a monomoraic vowel as a 'short' vowel.

3.4.4 Constraints

There exists a rather well established set of constraints for the treatment of stress within OT. Most constraints come from Prince and Smolensky (1993) and McCarthy and Prince (1993). We shall now suggest a set of constraints that can be used for the analysis of Swedish stress. We will base this both on the metrical analysis in Section 3.2.3 and the rule system for Swedish stress in Section 3.4.1. We will also include some general constraints, which follow the system used by Gussenhoven (2000), who analyses stress in Dutch. Further motivation of this particular choice of constraints is outside the scope of this chapter, but there seems to be a set of 'core constraints', which are generally assumed in the OT framework, from where we have selected the following set. For an alternative set of constraints as well as an alternative analysis of Swedish stress, see Shokri (2001).

An OT account of the metrical parameters

Let us repeat the metrical analysis of Swedish:

Boundedness:	YES
Quantity sensitivity:	YES
Foot headedness:	LEFT
Word headedness:	RIGHT
Directionality:	RIGHT-TO-LEFT

Transferring this to constraints in the OT framework, we get the following constraints:

- FOOT-BINARITY
  Feet consist of two syllables or two morae.
- WEIGHT-TO-STRESS PRINCIPLE (WSP)
  Bimoraic syllables are feet heads.
- RHYTHMTROCHEE
  Feet are left-headed.
- FOOTRIGHT (abbreviated F'RIGHT)
  Words are right-headed; the right edge of the word is aligned with the right edge of a strong foot.
- ALIGN-FOOT-RIGHT
  Feet are formed from right to left in the word

General constraints

The following constraints will also be used:

- GRWD=PRWD
  A grammatical word must be a prosodic word.
- STRESS-TO-WEIGHT PRINCIPLE (SWP)
  Foot heads are (minimally) bimoraic.
- SUPERHEAVY-TO-STRESS PRINCIPLE (SHSP)
  Trimoraic syllables are strong foot heads.
- NONFIN
  Primary stress does not appear on the final syllable.
- NOCLASH
  Foot heads are not adjacent.

- SylMon
  Syllables are monomoraic.

- Weight-by-Position' (WbP')
  Starting at the primary stressed syllable, coda consonants are moraic. In pre-stressed syllables, coda consonants are not moraic.

- HeadMatch(Ft)
  A foot head specified in the input form is also foot head in the output form.

Comments on the general constraints

The first constraint, GrWd=PrWd, demands that a grammatical word must have a foot. The prosodic hierarchy says that a prosodic word must have a foot as head, so the demand for a prosodic word implies a demand for a foot. The effect of this is to force at least one foot in the word, and hence forces monosyllabic words to have stress. This constraint has a high ranking and will be presupposed in the following analysis. Constraints WSP and SWP will be collapsed into one below, which will create the combined constraint that stressed syllables are heavy and heavy syllables are stressed. In mora terms this means that foot heads have at least two morae and that the weak syllable in a foot is monomoraic. SHSP will be used for final stress and is similar to WSP, but is a stricter version of it. RhythmTrochee is ranked high and will be taken for granted in the following discussion. This means that candidates with feet without an initial head (non-initial prominence) will be rejected without showing this in tableaux. The default pattern of penultimate stress is realized by NonFin, which forbids final stress, and F'Right, which imposes primary stress on the final foot. Together with RhythmTrochee this favours a final left-headed foot (=penultimate stress). NoClash prohibits two adjacent stressed syllables, which means that monosyllabic feet only occur word finally. The constraint Foot-binarity demands binary feet, either at mora or syllable level. Note that both monosyllabic and trimoraic feet are allowed. This constraint is highly ranked and will not be shown in all tableaux. The mora counting is treated by SylMon, WbP' and WSP. The default rule says: one syllable = one mora (SylMon) but the primary stressed syllable has at least two (WSP). From the primary stress and onward (post-stress position), coda consonants also count as morae (WbP'). Pre-stress coda consonants do not count. This is adopted from Gussenhoven (2000). This is important in final stressed words (see below). HeadMatch(Ft) takes care of exceptions.

3.4.5 Monomorphemic words

In Figure 3.2 the evaluation of /amøba/ is shown. The winning candidate satisfies all constraints. NOCLASH eliminates candidate b., and this makes the pre-stressed syllable extrametrical. Candidates c. and d. violate F'RIGHT, which forces the right edge of the strong foot to be aligned with the right edge of the word. Candidate e. has stress on the final syllable and it therefore rejected by NONFIN. Candidate f., finally, has a trisyllabic foot and violates FOOTBIN.

/amøba/	FOOTBIN	NOCLASH	NONFIN	WSP/SWP	F'RIGHT
☞ a. a'(mø:.ba)					
b. (a:)'(mø:.ba)		*!			
c. '(a:.mœ)ba					*!
d. '(a:.mœ)(ba:)					*!
e. (a:.mœ)'(ba:)			*!		
f. '(a:.mœ.ba)	*!				

Figure 3.2: Evaluation of /amøba/.

The analysis of words with closed penult, i.e. *gorilla* 'gorilla' gives the same result. The only difference is that the candidates corresponding to c. and d. in Figure 3.2 also will violate WSP/SWP, since they have a heavy (bimoraic) syllable in the weak position in the foot. In words with a closed antepenult, i.e. *armada* 'armada', the candidate corresponding to a. may be accused of violating WSP/SWP, since under a bimoraic analysis (vowel = one mora, consonant = one mora) the antepenult hasn't formed a foot. The bimoraic candidate will, however, also violate WBP'. If the consonant is not counted as a mora (satisfying WBP') the pre-stressed syllable does not form a foot, which gives the same result as in Figure 3.2. See Figure 3.3.

/armada/	NOCLASH	WSP/SWP	F'RIGHT	WBP'
☞ a. μ ar'(ma:.da)				
b. μμ ar'(ma:.da)		*!		*
c. (a:r)'(ma:.da)	*!			*
d. '(a:r.ma)da			*!	
e. '(a:r.ma)(da:)			*!	

Figure 3.3: Evaluation of /armada/.

Note that there are exceptions, like *ättika* 'vinegar' and *paprika* 'paprika', which have initial stress. These words are analysed as having a prespecified foot pattern, which forces stress on the correct syllable by adding a highly ranked constraint that demands that the foot structure in the input form and the winning candidate must be the same, i.e., HEADMATCH(FT). This is further discussed in Section 3.4.5.

In words with a closed final syllable the interaction between WSP/SWP and F'RIGHT is important. By ranking WSP/SWP higher, the right candidate emerges as winner both with open and closed penults. The analysis is shown in Figure 3.4. Candidate 1.a. violates F'RIGHT but not WSP/SWP; both foot heads are bimoraic and the only weak syllable is monomoraic. The competing candidates are ruled out through violations of more highly ranked constraints. A closed penult, however, causes a violation of WSP/SWP regardless of where primary stress is (2a.-d.). This means that F'RIGHT decides. One of the advantages with OT is evident here: the winning candidate may violate constraints as long as no other remaining candidate performs better.

/maraton/	NOCLASH	NONFIN	WSP/SWP	F'RIGHT
☞ 1a. '(mɑ:.ra)(tɔn)				*
1b. ma'(rɑ:.tɔn)			*!	
1c. (mɑ:)'(rɑ:.tɔn)	*!		*	

/reaktor/	NOCLASH	NONFIN	WSP/SWP	F'RIGHT
2a. '(re:.ak)(tɔr)			*	*!
☞ 2b. rɛ'(ak.tɔr)			*	
2c. (re:)'(ak.tɔr)	*!		*	
2d. '(re:.ak)tɔr			**!	
2e. (re:)'(ak)(tɔr)	*!*			*

Figure 3.4: Evaluation of /maraton/ and /reaktor/.

In bisyllabic words with closed final syllables and penultimate stress we see that NOCLASH and NONFIN are ranked higher than WSP/SWP, meaning that it is more important to avoid stress clash and final stress than that heavy syllables are unstressed. This is shown in Figure 3.5.

It remains to show how final stress is realized. Monosyllabic words are handled by GRWD=PRWD, but in polysyllabic words candidates with final stress are ruled out by NONFIN. However, by using a prespecified superheavy syllable in the input form and ranking SHSP higher than NONFIN, finally stressed candidates may be

/anis/	NoClash	NonFin	WSP/SWP	F'Right
☞ a. ’(ɑ:.nIs)			*	
b. a.’(nIs)		*!		
c. ’(ɑ:).(nIs)	*!			*
d. (ɑ:).’(nIs)	*!	*		

Figure 3.5: Evaluation of /anis/.

winners, see Figure 3.6. Note also that the ranking SylMon >> WSP/SWP rules out candidate d., since this has more bimoraic syllables (looking back at Figure 3.5 we can also establish the ranking NonFin >> SylMon, since otherwise candidate b. would have won the evaluation of /anis/). This also causes the foot to become a syllabic trochee, cf. the foot inventory in Section 3.2.4.

The reason that the superheavies must be specified is that *both* superheavies and normal heavy syllables can occur in final position, e.g., compare *anis* with *polis* ’police’, where the latter has final stress. Only the superheavy one receives stress. Therefore they must be specified in the input form. However, many of them occur in syllables that seem to be ’submorphemic’, e.g., that they are affixes etymologically. It is possible that they can be listed so as to reduce the number of words where it is necessary to prespecify a superheavy syllable (see also Section 3.4.5).

/katastro:f/	NoClash	SHSP	NonFin	SylMon	WSP/SWP	F'Right
☞ a. (ka.ta).’(stro:f)			*	*	*	
b. ’(kɑ:.ta).(stro:f)		*!		**		*
c. ka’(tɑ:.stro:f)		*!		**	*	
d. (kɑ:.ta).’(stro:f)			*	**!		

Figure 3.6: Evaluation of /katastro:f/.

Problems

There are a few problems in the present analysis. We have not motivated why we regard penultimate stress as the best default rule. Formally, the default rule could as well predict final stress, and penultimate stress marked in the lexicon. This is the approach followed by Shokri (2001) for Swedish. We take some support from similar analyses of Dutch (Gussenhoven, 2000) and German (Féry, 1998), where the penult has been used as default pattern, and claim that the similarities within the Germanic language family support this analysis for Swedish. There are also a

larger number of monomorphemic words with penultimate stress than with final or antepenultimate stress. The number of types of monomorphemes is perhaps not the most relevant factor; the daily use (the number of occurrences of tokens of each word) is probably more important. Another reason is that it is easier to find patterns (submorphemic similarities) at the end of a word than in the middle. Some of the words with a superheavy final syllable can be identified by their final syllable, e.g., the syllables *-åb*, *-ad*, *-id*, *-age* (orthographic representation) are all stressed.

The same case can be made for the prespecified foot patterns that were used in trisyllabic words. Whenever there is variation in the data, one must determine a default rule and then use exceptions, exceptions from the exceptions etc. We believe that the default patterns we have chosen cover a lot of cases, and that the words that are treated as exceptions in many cases can be handled by other linguistic rules, based on, e.g. submorphemic patterns, and schwa vowels (which never carry stress).

A third problem is that we may have followed Gussenhoven's argumentation for non-moraic pre-stress coda consonants too rigidly. Riad (1992) states that it has more to do with the degree of sonority of the consonant than the position in the word and Swedish and Dutch may differ in this respect. But this is an empirical question that we will leave unsolved at present. The constraint that realizes this phenomenon (WBP') has, after all, a low rank and is thus less important.

3.4.6 Compound words

Compound words receive a special stress pattern in Swedish. They often consist of two or more words or morphemes with one true word stress each, but only one of them is realized with primary stress in a compound word. Primary stress goes to the leftmost stressable element, while the last stressable element receives secondary stress. Other stresses (in compounds with more than two parts) are not realized, but they may affect the rhythmical pattern of the syllables between primary and secondary stress. In order to treat this, the clitic group level is used. When two prosodic words form a grammatical unit on this level, each prosodic word projects its own head, but since they are grouped together only one of these heads is realized with primary stress. By adding the constraint MAIN-LEFT(C), which says that the head of a clitic group should be to the left, primary stress on the first part of the compound is realized.

- MAIN-LEFT(C)
  A clitic group (C) is left-headed

As long as there is only one prosodic word, the head of the clitic group ends up on the same unit as the head of the word (CCC in Section 3.2.2). This makes the last

foot carry primary stress. When two or more words are grouped together in a clitic group, the head of the group will be in the domain of the first prosodic word, which gives primary stress to the last foot of the first prosodic word. Compare Tables 3.3 and 3.4. Note that each prosodic word first produces a head, then the position of primary and secondary stress is determined.

Table 3.3: Two one word phrases (two prosodic words, *maskin* 'machine' and *fonetik* 'phonetics').

C	(			x	)	(				x	)			
PWd	(			x	)	(				x	)			
Ft		<σ>	(	x	)	(	x	.	)	(	x	)		
		ma		ˈɧi:n			fʊ	nə			ˈti:k			

Table 3.4: One compound phrase (one prosodic word, *maskinfonetik* 'machine phonetics', opposed to ear phonetics).

C	(			x										)
PWd	(			x	)	(				x	)			
Ft		<σ>	(	x	)	(	x	.	)	(	x	)		
		ma		ˈɧi:n			fʊ	nə			ˌti:k			

3.4.7 Affixes

The influence of affixation on stress was mentioned in 3.4.1. The affixes follow the weight sensitivity restrictions mentioned above. Some affixes do not influence the stress pattern at all, since they do not contain any heavy syllable. Other affixes introduce a new heavy syllable, in addition to the one in the root of the word. This means that the resulting word contains two or more heavy syllables. The stress pattern of the word then follows the rules for compound words (see the previous section). Depending on whether the affix is a prefix or a suffix it will get primary stress or secondary stress.

Another group of affixes has a heavy syllable, and deprives the root of stress. These suffixes cause a change in the morphophonology of the root, i.e., a heavy syllable in the root becomes light, see Table 3.5. A way of analysing this is to assume that these suffixes have prespecified information that says that a syllable in the suffix must be the head in the clitic group. The other syllables have to adjust to this. Since only one heavy syllable remains, only one prosodic word is formed. Therefore, word stress will appear on the heavy syllable in the suffix.

It should be emphasised that the interaction between morphology and prosody is a lot more complex than presented here. Our main purpose is to show the

Table 3.5: Stress affected by a morpheme with a prespecified head at the C level.

C	(	x		)		(			x		)			
PWd	(	x		)		(			x		)			
Ft	(	x	.	)	<σ>	(	x	.	)	(	x	.	)	
		lɛ:	ra		rə		lɛ	ra			ri:n	a		

necessity of using the clitic group in treating the stress pattern in compounds and in morphologically complex words, since prosodic words have right-bounded primary stress, while compounds have left-bounded primary stress. This is an area where a more extensive analysis is needed.

3.5 Conclusions

We have shown that the phonology of Swedish stress is handled well within optimality theory using a correct constraint ranking hierarchy. The metrical analysis provides a useful starting point for the OT analysis, and this is extended with mora and syllable structure information. Swedish shows some variation from what the rules predict, and in the present analysis we handle this by assuming lexical features in the deviating words. The most important features of monomorphemic words is that a grammatical word must have at least one foot, and that there is a preference for left-headed, binary feet finally in the word. The basic foot type is the moraic trochee, but in pre-stress position syllabic trochees may occur. Final stress is avoided, unless marked in the lexicon. In order to realize this we have established the following constraint rankings:

1. RHYTHMTROCHEE, FTBIN, GRWD=PRWD, MAIN-LEFT(C): highest rank
2. {NOCLASH, NONFIN} >> WSP/SWP >> F'RIGHT
3. SHSP >> WSP/SWP
4. NONFIN >> SYLMON >> WSP/SWP

Complex words (compound words and derivatives with more than one morpheme) must be treated on a higher level than the prosodic word, since monomorphemic words prefer stress to the right in the word, whereas compound words are left-headed.

Chapter 4

Letter-to-sound: allophones and prosody

4.1 Introduction

In this chapter we develop a system for prediction of allophones and lexical prosody for unknown words in Swedish in a TTS context. Unknown words are words whose analysis and pronunciation can not be listed in a pronunciation dictionary, since these dictionaries can not contain an infinite number of words. This is problematic for Swedish, since it is a language that has many productive word formation processes, which can result in infinitely many words. The system enables the automatic assignment of phonetic segments, position of stress, and type of word accent based on Swedish orthography. We use a large lexicon with orthographic and phonetic transcriptions in order to build a statistical prediction model. The model uses different features of the orthographic transcription and is used to predict pronunciation in the form of phonetic segments as well as prosodic features such as position of primary stress and secondary stress as well as word accent. The model assumes no knowledge of morphological information, implying that it works for both simple and complex (compound) words. Prediction from both letter-by-letter and whole-word patterns is tested. Furthermore, the possibility of using a part-of-speech (POS) tag to improve the performance is examined.

4.1.1 Outline of the chapter

We start out with a description of the problem at hand, and comment on the different strategies that are used. Some extra attention is given to prosodic problems. Then we survey some of the earlier solutions. This survey concentrates on solutions for English, rather than Swedish, since English is the language that has been examined most frequently and for which most solutions are developed. This is followed by a description of the methodology chosen and then we move on to the

actual development and testing of the rules for Swedish.

4.2 The problem: from letter to sound

In the context of text-to-speech, we sooner or later face the problem of converting the units of texts (letters and other graphemes) to the units of speech (phones or allophones and prosody). In Chapter 2, we sketched the task of the pre-processing stage, where symbolic, junctural and numeric characters are treated. The problem studied here is how to go from the alphabetic level to the phonetic by only using orthographic information.

4.2.1 Dictionary-based and rule-based strategies

The problem is often tackled using a combination of *dictonary-based* and *rule-based* strategies. In dictionary-based approaches the pronunciation of a particular word is looked up in a straightforward manner in a, possibly very large, list of words collected in a dictionary. Basically, a dictionary is a list of lexical entries. The lexical entries consist of at least two parts: the head word (or keyword, the item used to search the lexicon) and a representation of its pronunciation. As much phonetic information as possible is stored in this representation, possibly even syllable boundaries and lexical stress and/or accent. By matching the word currently under analysis against the head words in the dictionary the pronunciation part may be retrieved.

Rule-based approaches compress the competence that is encoded in dictionaries by capturing regularities between spelling and pronunciation in rules. These are often called letter-to-sound (LTS) rules. The rules work like a transducer, where the orthographic representation is used as input. A transducer, put very simply, maps between one set of symbols and another. The task of the LTS rules can thus be seen as realizing a transducer: the orthographic representation is transduced into a phonetic representation through the LTS rules. In Sproat (2000), the relation between the phonological and the orthographic representations is viewed as one of *licensing*. This means that particular linguistic elements *licence* the occurrence of orthographic elements. The task of the LTS rules then becomes to uncover this licensing.

Rule-based approaches can be further characterized as either *hand-made* or *machine learned*. Hand-made (also *expert* or *knowledge-based*) rules are the result of a skilled language expert that examines the spelling-to-pronunciation correspondence and constructs the rules manually. Machine learned strategies (also *trained*, *self-organizing*, *stochastic* or *data-driven*) solve the problem by using pattern matching techniques and artificial intelligence in combination with a given (usually

large) machine readable pronunciation dictionary, and infer the rules by finding generalizations and similar letter-to-sound mappings in the dictionary.

Dictionary-based approaches, at least in theory, will always return a correct result. However, they only work if there is an exact match between the input word and an entry in the dictionary. In practice they will always suffer from incomplete coverage since the potential number of different words is infinite. Some words will always be unknown and it will be necessary to treat these with rules anyway. Rule-based solutions, on the other hand, always return a result; their coverage is – or rather: can be made – complete. However, the result may not always be correct.

Another issue, treated by Pagel et al. (1998) is that a rule-based system may be used to reduce the number of words in the dictionary, thereby reducing the size and the amount of searching in the dictionary. This is achieved by applying the rules for each item in the lexicon. If the transcription hence produced by the rules is identical to the one found in the lexicon, the entry may be removed. Using this technique, they have achieved a reduction of 78% for the British English OALD dictionary, and 94% for the French BRULEX dictionary.

4.2.2 LTS approaches for Swedish

There have been some efforts at rule-based systems for Swedish. Apart from several (closed) commercial TTS systems (from Infovox, Telia and L & H RealSpeak) that undoubtedly have rule-based components, Carlson and Granström (1976), Jonsson (1986), Torstensson (2002) and Uneson (2002) have independently developed rule systems for Swedish. In addition, Gustafson (1996) used rules for the transcription of Swedish names and also provided an extensive discussion of different techniques. Apparently, all these systems use hand-crafted rules, whereas the present approach uses machine learning techniques.

For the dictionary-based method, there exists an approach for Swedish in the form of the CTH lexicon (Hedelin et al., 1987). Commercial systems from Telia, Infovox and ScanSoft (former L&H) are also likely to utilize dictionaries.

In this chapter we will develop a rule-based system for letter-to-sound and letter-to-prosody prediction. This should not be interpreted as implying that we prefer one method to the other. Rather, as we stated initially, the problem of converting text units into speech units is best handled by a combination of both approaches. If we are to make best possible use of a combined model, both components must be as good as possible.

Development issues for dictionary-based methods

Another reason that we chose to work on rule-based methods is that it simply is more interesting from a modelling perspective. In fact, there may not seem to be much room for improvements with the dictionary-based method, except that

adding entries to the dictionary will result in better coverage. However, there are a few possible areas of research.

The first such area is **standardization**. A key factor for a successful lexicon is that it encodes *relevant* and *consistent* information. When developing a lexicon it is important that some form of control is implemented so that this can be achieved. Level of encoding (phonetic, phonological, syllabic etc) and detail (should there be a full vowel, a reduced vowel or epenthesis in unstressed syllables?) must be kept as constant as possible, otherwise this may cause undesired pronunciation variations. This is especially true if the lexicon is developed by several different persons, which may be the case since lexicons are large and require significant efforts. Wolff et al. (2002) have developed a system for measuring the quality of pronunciation dictionaries.

The second issue is **dialect independency** of lexicons. Most present lexicons are developed specifically for one dialect. If one wishes to build a speech synthesizer for another dialect, either the existing lexicon has to be modified or a new one has to be developed. Recently, some efforts have been made (e.g, Fitt and Isard, 1999) towards a lexical specification (for English) that is dialect independent and that allows the core form to be mapped to different dialects.

The third point is the **search method**. This is mainly a computational concern. Often a technique called *binary search trees* is utilized. This technique orders the entries according to some property (e.g., alphabetic spelling). Properly balanced trees are very efficient as each comparison may result in reducing the number of items left to inspect by one-half. This means that in a lexicon with one million entries, twenty comparisons will always be enough, as $2^{20} > 1000000$. Balancing the trees properly may however be nontrivial, especially when deleting existing entries or inserting new ones.

There are probably other issues as well, e.g., multi-linguality and adaptation of the lexicon for different speech rates. Even though these issues may offer areas for possible improvements, dictionary-based approaches have not been further examined in the present study. Some of the issues for dictionaries discussed here are perhaps more important for speech recognition than for synthesis.

4.2.3 Prosody

A somewhat overlooked area in LTS processing is the issue of lexical prosody, namely the assignment of stress position, and for some languages, word accent type. For a language like Swedish, where compounding abounds and makes listing all possible words impossible, the assignment of stress by rule is unavoidable. Given the importance of correct stress for perception, a rule system that is capable of predicting stress from orthographic spelling must be rated as a very important component in a TTS system. Correct assignment of word accent is perhaps not of

the same importance and is also sometimes trivial to predict. Still, it has some use if a completely lexicon-free LTS system has to be used, e.g. in very small applications, or where hardware memory sets a limit.

Stress

Bruce (1993, 1998) describes a rule system for prediction of stress position in simple (non-compound) words. This rule system was introduced and exemplified in Section 3.4.1. This extremely simple set of rules works very well for simple words. It relies heavily on correct syllable analysis, but well-performing systems for this exist (Bannert, 1998). A more serious problem is that it is rather difficult to know, on the basis of orthography only, whether a given word is simple or complex. It may be possible to recognize compound words to some extent on the basis of the identification of phonotactically illegal clusters (Brodda, 1979). However, this method is not completely robust as illustrated in Table 4.1. A further complication is that there might be more than one possible interpretation of a complex word, for instance, the word *elvispar* may be both *el-vispar* 'electric [hand]mixer' and *elvis-par* 'pair of Elvises', resulting in different stress patterns. Furthermore, the rules say little about what to actually do with complex words unless their structure is known and one has access to information on where the boundaries between the different parts are.

An optimality-theoretic account of Swedish stress was presented in Chapter 3. However, no evaluation of this system was performed and it was based on a phonological input form, not an orthographic one.

Word accent

Bruce (1977, 1998) also provides a system for prediction of word accent type for

Table 4.1: Examples where word-internal or whole-word orthography is insufficient to determine whether a word is a compound or not.

Compound	Non-compound
vinglas 'wineglass'	*vinglas* 'stagger (pass.)'
drivis 'drift ice'	*novis* 'novice' , *dagis* 'daycare center'
komage 'stomach of a cow'	*fromage* '[cold] mousse'
limejuice 'lime juice'	*limerick*
modellera 'plasticine'	*modellera* 'model (vb)'
publik 'similar to a pub'[1]	*publik* 'audience'
förband 'warm-up band (mus.)'	*förband* 'bandage'

non-compound words. This utilizes information on stress position and morphological structure. An implementation of this system was provided by Touati (1989), where the user could interactively input a word and its phonological features and receive a synthesized waveform of the word with the correct word accent. As this phonological information may be unavailable in a TTS system, we would like to predict the accent type directly from the orthographic form of the word.

4.2.4 Language dependence

The actual balance between dictionaries and rules is often language dependent. In some languages, like Finnish or Turkish, the pronunciation follows the spelling quite closely. In these cases, a rule-based approach may account for most of the words of the language. For a language like English, the classic example of a language with a large distance between spelling and pronunciation, where the same letter or letter combination may be pronounced in many different ways, a dictionary-based approach is useful to a much larger extent. The situation for prosodic features is similar. In the Romance languages, the position of the stressed syllable is highly or completely predictable from the orthography. In some languages with fixed stress, e.g. Czech, the problem hardly exists. However in languages like Russian or Swedish the position of stress may be harder to predict.

4.2.5 Swedish

There are a few problems specific to Swedish that are worth mentioning. A problem that is common to all vowels is that they all have at least two pronunciations: one "long" and one "short" variant. Long vowels only occur in stressed syllables. In a stressed syllable the vowel is long if it is followed by at most one consonant. The vowel is short if it is followed by more than one consonant (within the same morpheme, cf. *svans* 'tail' and *svan-s* 'of a swan'). It is, however, not always obvious from the orthography which syllable is stressed. Neither is it obvious where the syllable or morpheme boundaries are. Stress is sometimes indicated, in words like *idé* 'idea' and *succé* 'success', but these are exceptions. Word accent is not encoded in Swedish either.

The grapheme <o>[2] adds another complication: it is used for both of the phonemes /u/ and /o/, both in long and short versions, as shown in Table 4.2.

Compounding causes other problems. There are many sounds which are spelled with multiletter combinations, like <ng> for [ŋ] and <tj> for [ɕ]. Usually these are quite predictable by rules, but compounding creates clusters which do not follow the rules. In *hötjuga* 'hayfork' (from *hö* 'hay' + *tjuga* 'fork') the graphemes <tj> has the pronunciation [ɕ], but in *matjord* 'topsoil' (from *mat* 'food' + *jord*

[2]The notation with angled brackets indicates letter strings.

Table 4.2: Pronunciations of the grapheme <o>.

u:	*mot* 'against', *mod* 'courage', *ton* 'tone, note'
ʊ	*pokal* 'cup', *november*
o:	*son* 'son', *kol* 'coal'
ɔ	*kom* 'come (imp.)', *ton* 'metric ton'

'soil') <tj> is pronounced [tj]. There are prosodic problems with this too, as it might be even more difficult to determine syllable boundaries and stress positions across compound boundaries. Finally, loanwords that have retained their original spelling, like *limejuice*, *gnocchi* and *champagne*, pose further difficulties since they do not follow the spelling conventions of Swedish at all. Such words of foreign origin may also retain sounds from their source languages. Eklund and Lindström (2001) analyses this problem and suggests the term *Xenophones* for these sounds, in addition to examining possible ways of handling these sounds in text-to-speech and automatic speech recognition.

4.3 Letter-to-sound: earlier solutions

The following is a review of dictionary-based and rule-based letter-to-sound conversion methods for English. The review is focuses on English rather than Swedish, since most studies and new developments in the field are on English. One of the first more elaborated systems for LTS conversion was MITalk, presented in Allen et al. (1987). They used a dictionary of 12000 "morphs" (prefixes, roots and suffixes) containing pronunciation and part-of-speech (POS) information. Incoming words are analysed and given a pronunciation in terms of these morphs. For words that fail to obtain an analysis through the morph lexicon (new words, misspellings), a rule system with rules for affix stripping, letter conversion and stress assignment is used. Individual results for the LTS rules are not reported, but in the Modified Rhyme Test, where comprehension of initial and final consonants is tested, the error rate is 6.9%, and the authors state that the intelligibility of the speech is very high.

As for the trained rule-based systems, one of the first systems was NETtalk (Sejnowski and Rosenberg, 1986). They used a neural network to produce phonetic transcriptions from graphemes. The transcriptions were encoded through articulatory features, stress levels and syllable boundaries. Their network used a context of three letters on each side of the central grapheme. Their result per phone was 92% correct, which would correspond to a much lower per word score. One drawback with their system was that it only enabled monocharacter-to-monophone

predictions. This makes the performance worse when there is not one phone per letter in a word, as is common.

In Coker et al. (1990), the dictionary-based approach was further extended. Apart from using an even larger lexicon than MITalk, the authors introduced a process of *analogical extension*. By using existing lexical entries they infer the pronunciation of unknown words. This leads to a coverage of 89.6% of word types and 99.9% of word tokens in a multi-million word corpus, excluding names. Non-covered words are passed on to a letter-to-sound rule system. For names, a perceptual evaluation involving one listener gives 90% "Good" judgements.

The rule-based methods were revitalized during the late 90's. Luk and Damper (1996) used Markov models which allowed them to use variable width chunks of letters and phonemes, thereby remedying one of the limitations with NETtalk. The models were trained using a lexicon with words and pronunciations. They report 93.7% phones correctly and from this infer 75% words correctly. However, inferences from the 'phones correct' figure in other papers (Black et al., 1998) are somewhat lower. Furthermore, Luk and Damper's system did not include stress assignment. Another approach was used in Daelemans and Van den Bosch (1996, 2001). They use a method called lexicon-based generalization. Essentially, they use a dictionary to build a *decision tree* (see Section 4.4), in which they order the attributes in the context of a letter according to how important they are calculated to be. Another feature of their approach is to store as little context as possible, given that the mapping between grapheme and phoneme is unambiguous for a given lexicon. The context of the rules is allowed to expand until an unambiguous mapping between grapheme+context and phone is obtained. Their results are 93.5% correctly classified phones and, including stress markers, 63.6% flawlessly converted words.

Black et al. (1998) presented another method for building generalized pronunciation rule systems from lists of words and pronunciations. The method can be used both for allophones and for prosodic features. The predictions are learned from a large list of words and pronunciations with an automatic learning technique called classification and regression trees (CART, Breiman et al. (1984); see also Section 4.4 below). They report 95.8% correctly predicted phones, and for words they get 74.56% correct, including stress level markers. The same method has also been used for French (Pagel et al., 1998), with results reaching mid-90% for both phones and words. Given that this method has the best performance to our knowledge, and also that the proper tools to follow this method are available, we decided to try this method for Swedish as well.

4.3.1 Hand-crafted rules

Klatt (1987) quotes the figures in Bernstein and Pisoni (1980): 85% at word level

with random sampling of a large dictionary and from this infers 97% correct at the phone level. However, Damper et al. (1999)[3] performed a comparison of hand-made and machine learned rule-based techniques and found that the latter outperformed the former significantly. They used a dictionary of 16280 words.

4.4 Classification and regression trees

CART is a method that is used to construct decision trees automatically from data. Decision trees are a ranked set of yes-no-questions. A CART is a statistical model, which can deal with incomplete data and multiple types of features both in input features and predicted features. In an LTS context, the input features would be, for instance, strings of graphemes, and the output features would be allophones and other phonetic data. The rules that are produced are relatively human-readable since they are formulated in terms of questions about the data. This is in contrast to, e.g. the representation of knowledge in neural networks, where the complex interaction between nodes and their numeric weights makes them harder to understand. In this study, and in all the studies in this thesis, we used an implementation from the Edinburgh Speech Tools package (Taylor et al., 1999), which is called *wagon*. An illustration of a CART tree is shown in Section 8.3.2.

4.5 Automatic construction of rules

When building LTS rules it is common to use a lexicon where words and their pronunciations are listed. The idea is to try to find generalizations of the relation between word and pronunciation and to capture these in rules that can be applied to any unknown word for which we need a pronunciation.

Such a lexicon might contain many different kinds of information. The most important parts, however, are the word itself and its phonetic representation. We will call them the *orthographic* and the *allophonic* parts of the lexicon, respectively. Lexicons of this kind are usually allophonic, rather than phonemic, since phonemes would have to have a pronunciation anyway. Other types of information include part of speech, higher-level information such as stress or word accent and possibly morphological composition and etymological origin.

In Black et al. (1998), a method for learning LTS rules automatically is presented. This method can be described as 'alignment and rule induction'. With this method, rules are built automatically from a lexicon. After alignment between words and pronunciations (described below), feature vectors are built for each letter for all words in the lexicon. The feature vectors contain the letter, the allophone that

[3] Damper et al. (1999, p. 7) report that they are unable to find this actual claim in the paper but that Bernstein has confirmed the figure in personal communication.

should be produced by that letter, and the context letters, which help in ambiguous cases, where one and the same letter might give different allophones because of irregularities in the spelling-to-pronunciation system. Other information such as part of speech, syllable weight and morphology might also be included if it is believed to help. See Table 4.3 for an example.

Feature vectors are then sorted and grouped so that all feature vectors containing the same predictor letter are put in the same group. This will result in feature vectors with different context constellations and (in most cases) different allophones. Part of such a group is shown in Table 4.4.

The CART procedure then examines all the feature vectors for a given letter, and by calculating the frequency distributions – and from them, the entropy impurities (see below) – of the instances of the different attributes of the feature vectors a structured set of yes-no questions is formulated in the form of a tree. This tree will ultimately be the prediction procedure for that letter. The procedure examines all examples in the data set and selects those that give the best information, i.e.

Table 4.3: Feature vectors for the word *ackordlöns* '[of] piece wages' The top row contains the spelling, the second row the allophones, as pronounced in Standard Swedish. The underscores (_) are empty slots in the allophonic representation (See Section 4.5.2). Below these we see the feature vectors, one on each line. The columns labelled '1' and '2' contain the allophones and the predictor letters, respectively. The other columns show the context of each letter. Word boundary is denoted by # and zeroes fill out the positions beyond the word boundary.

a	c	k	o	r	d	l	ö	n	s
a	_	k	o:	_	ɖ	ɭ	ø:	n	s
1					2				
a	0	0	0	#	a	c	k	o	r
_	0	0	#	a	c	k	o	r	d
k	0	#	a	c	k	o	r	d	l
o:	#	a	c	k	o	r	d	l	ö
_	a	c	k	o	r	d	l	ö	n
ɖ	c	k	o	r	d	l	ö	n	s
ɭ	k	o	r	d	l	ö	n	s	#
ø:	o	r	d	l	ö	n	s	#	0
n	r	d	l	ö	n	s	#	0	0
s	d	l	ö	n	s	#	0	0	0

it automatically determines which aspects of a letter's surroundings are the most revealing in order to determine how that letter should be realized. When predicting a pronunciation for an unknown word, an allophone is determined by feeding each letter (with its context) as a feature vector into the appropriate tree. The classification is done by sorting the letter down the tree from the root node to some terminal leaf node, testing the attributes of the feature vector at the different locations in the tree. The actual allophone classification is provided by the contents of the leaf node that is ultimately reached. All allophones are then concatenated to produce the full pronunciation of the word.

4.5.1 Impurity and entropy

When building trees, the idea is to examine a set of feature vectors and, from them, create subsets that are as *pure* as possible, i.e. where the samples are as similar to each other as possible. This is the same as decreasing the *impurity* of the set of samples. In the *wagon* program, the impurity of sample sets with discrete predictees – like

Table 4.4: Grouped feature vectors. The columns labelled '1' and '2' contain the allophones and the predictor letters, respectively. Note the different realisations [a] and [ɑ:] of the grapheme <a> depending on the letter context. Note that the word *abakus* 'abacus' may also be pronounced with the first vowel as long. The transcription is taken from the CTH lexicon (Hedelin et al., 1987).

1					2				
a	0	0	0	#	a	b	a	k	u
a	0	#	a	b	a	k	u	s	#
a	0	0	0	#	a	b	a	n	d
a	0	#	a	b	a	n	d	o	n
a	0	0	0	#	a	b	b	e	d
a	s	o	r	n	a	#	0	0	0
a	0	0	0	#	a	b	b	o	r
ɑ:	b	o	r	r	a	r	t	a	d
a	r	a	r	t	a	d	#	0	0
a	0	0	0	#	a	b	b	o	r
ɑ:	b	o	r	r	a	r	t	e	r
a	t	e	r	n	a	s	#	0	0
a	0	0	0	#	a	b	b	o	r

allophones – is measured by the entropy[4] of the sample set[5]. Entropy is calculated by the formula:

```
-1 * (sumof for each x in class prob(x)*log(prob(x)))
```

Let us give an example. Assume that the data for the grapheme <g> has the following allophonic distribution[6]:

- g 56%
- ŋ 44%

When calculating entropy, the percentages are converted to probabilities or fractions, meaning that 56% becomes 0.56. In this case the 'class' is the letter <g> and the different 'x':s are the instances of this class: the allophones g and ŋ. These instances have the probabilities that we derived above. The entropy thus is:

```
-1 * (0.56 * log(0.56) + 0.44 * log(0.44)) = 0.6859.
```

Now, we formulate a question that will be tested. The question is: Is the preceding letter an <n>? The preceding letter is <n> in 44% of the cases. The distribution of realizations of <g> if the preceding letter is <n> is:

- g 2%
- ŋ 98%

Entropy:

```
-1 * (0.02 * log(0.02) + 0.98 * log(0.98)) = 0.098
```

The preceding letter is something else than <n> in 56% of the cases. The distribution of realizations of <g> if the preceding letter is something else is:

- g 97%
- ŋ 3%

Entropy:

[4]The concept of entropy is widely used in information theory, where it is a measure of the number of bits needed to encode an arbitrary member of the sample set.

[5]Actually, it is also weighted by the number of samples. This favours larger partitions, leading to better trees.

[6]We are ignoring other possibilities here like j, k and ʃ – these would increase the entropy.

```
-1 * (0.97 * log(0.97) + 0.03 * log(0.03)) = 0.1347
```

The total entropy is:

```
0.44 * 0.098 + 0.56 * 0.1347 = 0.1186
```

This is a huge decrease from 0.6859, which means that the impurity has decreased and that this question probably is a good one. When looking for questions, entropies are calculated in this way and the question that results in the lowest entropy is selected, as this leads to the purest possible subsets. The best question is then added to the tree, and since questions are of type yes/no, each new question effectively splits the data into two parts. The process is then applied recursively on newly formed subsets until some stop criterion, such as a minimum number of samples (feature vectors) in a group, is reached.

4.5.2 Aligning letters and allophones

This method of building LTS rules requires pre-aligning the orthographic part and the allophonic part of the words in the lexicon. This is because we need to know which letter corresponds to which allophone. This is usually not given in the lexicon since it simply lists the whole word as a letter string and the pronunciation as an allophone string. Spelling with double consonants (like the <ll> in 'spelling'), and different multiletter spellings of sounds that "do not have a letter" then causes the letters and the sounds that actually do have a relationship to be positionally displaced. This is evident as the number of letters and the number of sounds often differ, as shown in Table 4.5. The number of letters is usually higher. Thus, the alignment is not a one-to-one-mapping since every single letter in a word does not correspond to a single allophone.

The solution proposed by Black et al. (1998) is to insert empty slots in the allophonic representation. These slots are called _epsilon_ and we will refer to

Table 4.5: Unaligned and aligned sequences of letters and allophones. Underscore characters denote empty slots.

Unaligned				
t	j	o	c	k
ɕ	ɔ	k	_	_
Aligned				
t	j	o	c	k
ɕ	_	ɔ	_	k

them either with the underscore character '_' or with the word "_epsilon_". The alignment procedure might be performed either by hand or automatically. Hand-made alignments would be the best as they can always be made to represent the correct mapping between letters and allophones. However, the amount of training data usually needed makes it impossible to hand-align the whole lexicon. Black et al. (1998) and also Daelemans and Van den Bosch (1996) present techniques for automatically aligning the training data. Black et al. (1998) also have a method where a certain amount of manual work facilitates the automatic method and they actually report a higher number of words where the complete phone string is correctly predicted when using this method.

Note that aligning letters and allophones in this way is a much easier task than aligning acoustic speech frames, as is done for the normalization of different speaking rates and for automatic labelling (see Section 8.4.1). Since both alphabets are discrete, much simpler distance measures may be used.

4.6 Experiments

Here we will describe our experiments. We have attempted to create LTS rules both for the segmental (allophonic) transcription and for the generation of lexical prosody.

4.6.1 Segmental transcription

We followed the method of Black et al. (1998) to build LTS rules for Swedish. The CTH lexicon (Hedelin et al., 1987) was used but needed some processing. We removed all words from closed word classes – keeping nouns, verbs and adjectives. We also removed all abbreviations, and all entries beginning with a capital, thus excluding place names as well as proper names.

Lexicon expansion

The CTH lexicon lists uninflected word forms only. Since the system of rule extraction captures generalizations from its input data, and we want to produce rules that can be applied to arbitrary texts, we need to include inflected forms in the input data as well. Word endings will otherwise be treated by rules based on uninflected forms only, which will most probably lead to erroneous results.

In order to produce inflected forms, we use a procedure of *lexicon expansion* that follows Jonsson (1986). The expansion rules use the part of speech tag assigned to each word in the dictionary, as well as a key which describes the last characters of a word. The base forms, which are listed in the dictionary with orthographic and phonetic transcriptions, are then associated with a given inflection paradigm that

specifies which forms should be generated, as well as the additions and modifications necessary to produce these forms.

The expansion rules may overgenerate. This is, however, not a major problem since we will not use the expanded lexicon directly, but indirectly, as the basis for rule extraction. A few non-existing words do not harm the performance of the rules as long as they are consistent with the phonology in general.

There is no overt measurement of correctness mentioned in Jonsson's study, only a statement (ibid, p. 3-20) that:

> "The resulting phonetic transcription is of high quality and includes tonal accents and stress located in the correct syllables"

Regarding the coverage of the rules, it is reported that 95% of the words in unknown texts are found and transcribed.

Training and test sets

After expansion of the CTH dictionary the number of words totalled 781251. Since this requires an incredible amount of processing, we used only 10% of the words. These words were extracted by taking every tenth word from an alphabetically sorted version of the full list.[7]

This extracted list contained 78125 words. This new word list was in turn split into a training set and a test set, using 10% of the words for testing (70313 and 7812 words, respectively). The sets were subsequently processed so that the CART-building program (*wagon*) could process them. This included extracting letter context information, removing prosodic information where appropriate, and converting the transcriptions to the SAMPA alphabet.

Aligning

Since the relation between the letters and the pronunciation of a word is not one-to-one, the data needs to be aligned. We used the alignment procedure described in Black et al. (1998)[8]. Unless one is extremely meticulous about the alignment, some words will always be discarded since no alignment can be found[9]. After a reasonable amount of work, 70043 (99.62%) of the words in the training set and 7777 (99.55%) of the words in the test set were aligned correctly.

[7] Since the list was sorted alphabetically and since the number of morphological derivations for each word is irregular, we estimated that this procedure was adequate for producing enough spelling-to-pronunciation phenomena and a reliable mix of word classes, endings, and letters.

[8] A step-by-step description is provided in Black and Lenzo (2003).

[9] Many of these are loanwords of foreign origin, like *beagle* and *borsjtj*, and/or words that have more than one multiletter-to-allophone relation, like *diftongljud* 'diphthong sound' [diftɔŋjʉ̟d].

Feature extraction

We used a context of four letters before and four letters after the target letter. In this first run, we did not include part-of-speech tags in the feature set.

Model building

Models were then built, first separately for each letter and then combined into one big tree. Two runs were made, where we varied the "stop" value[10], setting it to either 1 or 2. The lower the value, the more fine-tuned to the training set the models get and there is a risk that the models get over-trained. However, Black et al. (1998) report that a value of 1 still gives the best results.

Results

The results per letter and per word are shown in Table 4.6. Also shown is the size of the resulting tree, counted as the number of nodes the tree contains. In effect, this is the number of rules in the system.

Table 4.6: Results per letter and per word for different stop conditions.

	Correct		
Stop	Letters	Words	Size
2	96.37%	68.61%	32877
1	96.87%	72.26%	42441

For the best tree (where Stop = 1), the individual results for each letter are shown in Table 4.7

Table 4.7: Results per letter.

Letter	Number	Correct	Percent
a	8509	8198	96.35%
b	1600	1598	99.88%
c	665	653	98.20%
d	3223	3112	96.56%
e	8060	7606	94.37%
é	33	33	100.00%
è	0	0	NaN

continued on next page

[10] According to Black et al. (1998) this value "specifies the minimum number of examples necessary in the training set before a question is hypothesized to distinguish the group".

continued from previous page			
Letter	Number	Correct	Percent
f	1684	1679	99.70%
g	3539	3365	95.08%
h	954	946	99.16%
i	5001	4718	94.34%
j	657	638	97.11%
k	3654	3472	95.02%
l	4701	4648	98.87%
m	2209	2174	98.42%
n	8612	8474	98.40%
o	3290	2920	88.75%
p	1955	1945	99.49%
q	3	3	100.00%
r	8438	8320	98.60%
s	9816	9588	97.68%
t	6769	6723	99.32%
u	1878	1782	94.89%
v	1668	1643	98.50%
w	10	9	90.00%
x	157	157	100.00%
y	696	642	92.24%
z	18	18	100.00%
å	703	678	96.44%
ä	1654	1606	97.10%
ö	1133	1082	95.50%
Total	91289	88430	96.87%
All vowels	30957	29265	94.53%
All consonants	60332	59165	98.07%

Note that the problems we anticipated for the letter <o> in Section 4.2.5 are manifested here; it has the lowest score of all letters that have high occurrence frequencies. The best-performing letter is <x> (and with much lower frequency of occurrence <é>, <q> and <z>). <b>, <f>, <p> , <t> and <h> also score above 99%. The consonant letters outperform the vowel letters, which was also anticipated since the long/short distinction for vowels is not coded in the orthography. An analysis of the errors showed that 54% of the errors for vowels (32% of all errors at the phone level) were errors in vowel length. Similarly, 6.4% (3.8%) were associated with the letter <o>.

4.6.2 Comparisons

Results for other languages were reported in Section 4.3. Torstensson (2002), using a knowledge-based technique for Swedish, reaches around 98% phones correct and 78% words correct, without prosody markers, with a test set of around 3500 unique words. These are admittedly high figures and a good result. However, one should be careful in comparing his study with the present one as different test materials are used. For instance, our material consists of content words only. We also suspect that our material might contain more compounds (and hence problematic junctures), as the word types in his material (excerpts from a novel + rock lyrics) follow the frequency distribution of texts, where non-compound word types are more frequent than compound word types. Furthermore, it is not clear in Torstensson (2002) what the relation is between, on the one hand, training/development material, and on the other, test material. The development of rules is described as being an iterative process, where elaboration of rules and testing is performed interchangeably. In theory, as soon as you change your rule system in order to correct some error discovered in a test material, a new test set must be used if you want your results to reflect the performance of truly unknown words. Damper et al. (1999, p. 160) writes:

> "The problem for TTS systems is rare/unusual words which cannot be anticipated."

The results reported here are optained with an absolute distinction between training material and testing material. If we, however, test the rules on the same material that we trained it on, our method scores 98.6% on the phone level and 87.5% on the word level for a material containing 70043 words.

4.6.3 Prosody prediction by letter

Prediction of stress position only

The first attempt at building rules for stress assignment used a mapping of vowel letters to either 1 (stressed) or 0 (unstressed), while all consonant letters were mapped to C. Note that these category names are rather arbitrary. Primary and secondary stress were treated equally. Secondary stress typically comes after primary stress, so it should be possible to predict it accurately from among the resulting stressed syllables. However, in this test, no limits on the number of stresses in a word were imposed. In theory, with the method used, stress could be predicted for every vowel letter in a word, or, indeed, for none of the vowel letters. This obviously renders incorrect word stress patterns, as shown in Table 4.8.

The same word list, training/test set split, alignment methods and feature extraction were used here, as were used for the segmental prediction. Models were

Table 4.8: Incorrect stress prediction of the word *analfabet* 'illiterate', where stress level 1 fails to be predicted for any of the vowels.

Word	#	a	n	a	l	f	a	b	e	t	#
Correct	#	0	C	0	C	C	0	C	1	C	#
Predicted	#	0	C	0	C	C	0	C	0	C	#

built for vowel letters only, as consonant letters always were mapped to the symbol C. The results under these conditions are: correctly predicted stress levels for all vowel letters is 97.03%, correctly predicted words is 73.25%. For individual vowel letters, the number ranges from 90.35% (worst case) for the letter <y> to 95.09% (best case) for the letter <å>. This means that the positions of stressed syllables are correctly predicted for 73.25% of the words in the test set. An analysis of the predicted stress patterns showed that impossible/illicit stress patterns were produced in 7.7% of the test words.

Full prosody prediction: position of stresses, main and secondary stress, and word accent type

Our next attempt involved using a more elaborate prosody model. Still using the same training and test set, we now used the following four categories when training the trees for the vowels:

0 - unstressed

2 - secondary stress

3 - word accent 2 (grave)

4 - word accent 1 (acute)

In the test, one of these levels was predicted for each vowel letter in a word. This means that we predict the position of stress (this is indicated by levels 2-4, unstressed positions have 0), the distinction between main (3 and 4) and secondary stress (2), and the distinction between word accent type (3 or 4). We had no restrictions regarding the number of stresses and word accents. This model gives 96.75% correct per vowel letter and 68.94% correct per word. The best vowel letter was <a>, with 91.99% correct and the worst was <y>, with 85.52%. An analysis of the predicted accent patterns showed that impossible/illicit accent patterns were produced in 21.3% of the test words.

Combination of allophonic and prosodic prediction

In the next test, we combined the letter and prosody prediction. Both the allophonic content and (for the vowels) the prosodic level was predicted. The results per letter were still quite good (94.61%), whereas the per word score drops to 55.24% correct. The best vowel was the letter <å>, with 92.32% correct and the worst was the letter <y>, with 77.01%. An analysis of the predicted accent patterns showed that impossible/illicit accent patterns were produced in 22.6% of the test words.

Adding part of speech information

We also experimented with adding part of speech (POS) information. Black et al. (1998) report a significant improvement both for allophone prediction and stress prediction. However, in our study, adding POS information does not show any improvement of the phone prediction. This is contrary to Black et al. (1998), but a possible explanation for this is that for Swedish, much of the POS information is already visible in the letter context in the form of the grammatical ending of the word. For stress prediction, a small improvement is achieved when adding POS: phones 97.07% correct, words 73.56% correct. We did not attempt prediction of word accent type using POS, since the other results where quite discouraging.

Discussion

As we have said above, one of the reasons that the per word score is rather low is that the predictions made for the vowels are independent of the other vowels in the word. This means that each vowel letter could result in a stressed or accented vowel phone, while there is a real limit of two stresses per word, of which only one may be accented. It also implies that sometimes none of the vowel letters in the word gets stress or accent, whereas the minimum should be one accented vowel letter. These errors do not affect the per letter rate at all, but do significant harm to the per word rate. A method that limits the allowed number of stresses and accents per word could remedy this, and improve the word score.

One possible such method would be to combine the output of the trees with an OT-like analysis, like the one presented in Chapter 3. One could conceive of the decision tree outputs as a GEN component of an OT system. As the trees are designed now, their outputs consist only of the most probable candidate (the one with the highest frequency of occurrence) for a certain combination of attribute values. However, it is actually possible to retrieve the second most probable etc. Allowing this for all vowel letters would produce several alternative patterns that could be utilized as input candidates for an OT ranking system.

Another method that makes predictions based on the patterns of whole words is presented in the following subsection.

4.6.4 Prosody prediction by whole-word patterns

In the approach used so far, a prediction is made for all vowels in a word. With this positional prediction, we predict a certain feature for each position. This is suitable for phone prediction, since, after all, most letters in a word have a correspondence with the phones. For prosody, this method gives us the position of the stressed and accented syllables. However, since this method requires that we make a prediction for all the letters in a word, one value (stress or accent level) for each vowel is predicted, which means that more than one may be stressed.

For prosody it is possible to predict an overall pattern based on features of the whole word. This can be done by only making one prediction for the whole word using the whole word's features. We can not use the letter context surrounding a given letter, since we are not using the individual letters as predictors. Instead, we use the whole word and its features, such as number of letters, number of vowels, first letter, second letter, first vowel etc. The returned prediction is the position, counted either in letters or in vowel letters, of main or secondary stress. It is also possible to predict global features, such as the overall accent pattern in a word. However, in this case we do not get the position of the stresses and accents. For accent pattern we used four possible categories: Accent 1, Accent 2, Accent 1+Secondary Stress and Accent 2+Secondary Stress. We used the following word features: the first ten letters, the last ten letters, the first five vowel letters, the last five vowel letters, the total number of letters, the total number of vowel letters and a part-of-speech tag. The same training and test material as for the letter-by-letter prediction was used. Table 4.9 shows the results.

Table 4.9: Results for word prosodic predictions from whole words.

Prediction	Correct
Position of main stress	88.6%
Position of secondary stress	87.1%
Position of leftmost stress (= position of main stress)	88.6%
Position of rightmost stress	95.8%
Accent pattern	87.3%
Word accent	87.7%
Existence of secondary stress	88.3%

The combination of position of leftmost stress and position of rightmost stress, in effect, produces the correct stress pattern for the whole word. If we calculate the 'worst case' scenario, where there is absolutely no overlap between the erroneously predicted words in these two groups we still get 84.4% correct (88.6 - (100 - 95.8)

= 84.4). The actual figure is likely to be higher, as it is reasonable to believe that some word may fail in both cases. Using the same analysis for the word accent pattern, we get 72.1% correct (88.6 - (100 - 95.8) - (100-87.7) = 71.1). Again, the actual score is expected to be higher.

These results are clearly better than the prosodic results for letter-by-letter prediction and they could possibly be used together with the allophone prediction in order to produce better overall scores for allophone+prosody. This shows that prediction from whole-word features is better than letter-by-letter features for prosody. The advantage with this method is that it is not possible to produce any illicit/impossible patterns.

4.7 Final discussion

The basic idea behind letter-to-sound conversion for text-to-speech is to cover as many words as possible by rules and to combine these rules with a small dictionary that contains the words that are most common. Even if not all words are perfectly transcribed, a properly functioning rule-based system will at least be able to produce something that is reasonably correct. A dictionary-based approach, on the other hand, will always perform correctly – unfortunately, however, only for a limited number of words. For words not covered by the dictionary, rules must be used. If the rules are general enough, they will also be able to produce results that are identical to the items in the dictionary. These entries may then be removed, and the size of the lexicon reduced. Rule-based techniques then amount to a form of lexical modelling, where the knowledge encoded in the dictionary is instead expressed in the form of rules. We thus achieve a lexical modelling of prosody, both of word stress and word accents – and, through the allophones, of vowel length.

There are a few more issues we would like to touch upon here. The first is that these predictions are completely word-internal and do not model coarticulation effects across word boundaries. In a real text-to-speech synthesis system, this has to be achieved in order to increase the naturalness of synthesis. Existing solutions often use some form of post-lexical rules, where the produced transcription is fed into a small, often hand-crafted set of rules to take care of things like vowel epenthesis and supradentalization across word boundaries. However, we think that it would be possible to model this using a method like the one presented in this chapter. In that case, we should not use a dictionary to generate our modelling data but instead use phonetically segmented data obtained from real speech.

Another thing is that the size of the phone set possibly affects the results. If you must select between twenty allophones you are likely to get 5% correct by chance. If you have 50 allophones, you only have 2%. We have used a phone set with 50 allophones. However, each letter may only be mapped to a certain number of allophones, ranging from 11 different allophones for <e> and <u> to 1 allophone

for <h>, <m> and <q>.

The last point is that when calculating per word scores, a uniform scoring algorithm has been used. All errors are counted equally, whereas in reality some types of errors are not as serious as others, e.g. the mapping of a <v> to [f] or [v] might not be audible in synthesis, whereas it does make a great difference if <k> is mapped to [k] or [ɕ], or if the stress pattern is wrong. Non-uniform scoring algorithms and a perception test with speech synthesized from the produced transcriptions will be the ultimate test of these rules.

4.7.1 Improvements, further research

There are several possible avenues for further research in this field. Some form of morphological parser would possibly be useful. The identification of stress-attracting and stress-rejecting morphemes is likely to improve the results. A brute-force strategy is to choose a larger number of words from the original dictionary. This would increase the size of the training data, possibly improving the results. Another, more linguistically grounded strategy is to pair up the vowels and only predict the 'archivowel'. This would get the base vowel correct more often. For the prediction of vowel length perhaps the output of the stress prediction could be utilized.

4.8 Concluding summary

We have developed a trained rule-based letter-to-sound prediction system for Swedish. Using a large lexicon, rules were inferred using a machine learning technique. Using a letter-by-letter prediction technique, we obtain 96.87% correct when predicting allophones from letters. Whole word patterns, where all the allophones in a word are correct, are correctly predicted in 72.26% of the cases. For prosody, the scores on the word level are: 73.56% for stress patterns, 68.94% for word accent + stress pattern, and 55.24 for full allophone+prosody prediction.

Predictions based on whole-word patterns perform better for prosody. The position of main stress is correctly predicted in 88.6% of cases, whereas the full stress pattern gets a conjectured 'not-worse-than' score of around 84% and word accent+stress gets no worse than 72%.

We think that the scores for the features of stress and word accent show that it is quite possible to model lexical prosody from orthography through the use of modern LTS techniques. However, the most problematic factor for allophone prediction is also prosodic, namely the distinction between long and short vowels.

Chapter 5

A survey of the modelling of intonation for speech synthesis in Swedish

5.1 Introduction

In this chapter, we describe earlier work on Swedish intonation, with special emphasis on work aimed towards use in TTS. This chapter is motivated since, in the modelling in later chapters, Swedish is the target language, and this makes it appropriate to describe the efforts on Swedish intonation research here. Furthermore, summaries that relate intonation models to Swedish TTS are rather sparse. We will also discuss the models with respect to how they account for the relationship between local and global intonation patterns and their relationship with autosegmental phonology.

There are fundamentally two strands of models that have been used for Swedish TTS: one within the framework of autosegmental phonology and one within the superpositional view based on speech physiology.

5.2 The model of Carlson and Granström (1973)

One of the earliest attempts at the development of a prosodic component of a text-to-speech system for Swedish is Carlson and Granström (1973). In this study, the first steps toward a rule system for prosodic description of sentences are presented. The system's capabilities are described as:

> "a tentative system predicting the fundamental frequency pattern of Swedish nonsense word sentences without limitations concerning word accent" (ibid., p. 32)

In the system, an intonation contour is viewed as a combination of two components: *sentence intonation* and *word stress marking*. Sentence intonation is modelled as a simple linear fall (on a Hz scale) with constant start and end values. Stress marking is then added by using stress, accent and segment durations to predict the temporal location of F0 anchoring points. In acute (A1) words such points are called minima and in grave (A2) words they are called maxima. Minima are placed directly on the level of the sentence component whereas the level of maxima is calculated from the duration of the preceding word. Cosine waves are used to connect the points, thereby obtaining an F0 contour for the whole sentence. For two successive minima, the amplitude of the wave, and thereby the 'height' of the accent, is determined through a formula based on the duration of the preceding stressed vowel.

Even though this, admittedly tentative, early system lacks a few important features, most notably a focus component, it lay the ground for much of later intonation modelling in TTS. Several principles developed here are still used in systems today:

- the separation of *global* and *local* parts of an intonation contour: one sentence component and one word stress component
- the anchoring of accent patterns to the stressed syllable
- the modelling of accent as one complete gesture; no separation of rising and falling parts of the accent is made
- the use of linguistically rich input to drive the prediction

The lack of a focus component is perhaps best explained by the theoretical status in the field at the time. The separation of accent and focus and how to model it had not yet been clarified. The authors went on to develop the commercial Infovox TTS system, still one of the world's most successful speech synthesis systems.

5.3 The model of Bruce (1977) and its later developments

The model developed and presented by Bruce (1977) introduces many new ideas and should be viewed as theoretically independent from the Carlson-Granström model, even though some common aspects exist regarding internal representations, see below. It is also a rule-based model and according to the following quote, is consists of:

> "a set of preliminary rules for generating F0-contours of Swedish sentences. These rules concern primarily word accent, sentence accent and terminal juncture" (ibid., p. 129)

Regarding the coverage of the rules, the following is claimed:

> "the rules are valid not only for Stockholm Swedish but are applicable in essential parts to the Central Standard Swedish area in general" (ibid., p. 129)

The work is to some extent based on earlier work by Gårding and Lindblad (1973), but has several modifications and refinements.

Bruce's model consists of three types of rules: pitch rules, F0-rules and join rules. The input to the model is a phonetic transcription where stress, word accent, sentence accent and junctures are included. The pitch rules utilize this prosodic information in order to produce a series of tonal points, which represent the essential tonal characteristics of an utterance. The information included in a tonal point is its temporal location and its relative pitch, specified as either HIGH (H) or LOW (L). This representation thus resembles the minima and maxima in Carlson and Granström (1973). However, the rules for introducing them are more elaborate. There are two types of pitch rules, basic and context-dependent.

The basic rules are responsible for mapping the prosodic features in the transcription onto tonal points, whereas the task of the context-dependent rules is to realize the influence that adjacent tonal points have on each other. In the 1977 version of the model, temporal displacement of a tonal point is the only such effect. As we will see, later versions (see Section 5.3.5) will allow remapping an L as an H as well as deleting/inserting pitch points.

The rules for sentence accents and junctures specify one tonal point (either H or L), whereas the rules for word accents specify two points: one H and one L. The difference between A1 and A2 is assumed to be the temporal alignment of the tonal points relative to the stressed syllable. Thus, the H of an A1 word is placed in the pre-stress syllable, whereas the H of an A2 word is placed in the stressed syllable, with the L coming in the following syllable in both cases. The sentence accent rules place an H after the stressed syllable and the juncture rules place an L in the initial and final syllables.

The points are temporally specified with reference to the syllable. As a syllable may contain more than one point, there are conventions on how to place them. Within the syllable, the first point is placed in the initial part of the vowel. This is based on the assumption that vowel onsets are perceptually important for tonal events. When there are more than one point in the same syllable, the others are placed in the middle or at the end.

The F0-rules take care of the transition from relative pitch levels to actual F0-levels[1]. Each pitch point is mapped to one of the levels 1, 2, 3 or 4 (from lowest to highest) depending on their relative level (H or L), underlying prosodic category (word or sentence accent; or juncture) and the nature of adjacent points. Bruce, perhaps surprisingly, assumes a flat base line and uses this for level 1. The other levels come at increasingly higher positive distances from the base line. The join rules, finally, simply fill out the portions between the F0-points through linear interpolation. The 'new' features in this model are numerous. Here we simply list the most important ones:

- the anchoring of tonal movements to a perceptually relevant point in the syllable, the vowel onset
- the inclusion of an additional level of prominence through the sentence accents
- the inclusion of junctures
- the use of different pitch levels to account for the relationship between relative and actual F0-levels
- the pairing and simplification of the word accent distinction through the idea: same gesture, different timing

5.3.1 Downstepping

In Bruce (1982), downstepping in Swedish was analysed and incorporated in the model. Downstepping is seen as triggered by the occurrence of a sentence – or *focal* – accent. After a focal accent, the successive non-focal accents are downstepped. Downstepping is an expression of the equal prominence of the post-focal accents within the phrase. Downstepping is modelled by a successive lowering of the F0 levels of the post-focal accents.

5.3.2 Phrasing

Work towards the modelling of phrasing in TTS is reported in Bruce and Granström (1990). In this paper, phrase-internal boundaries are realized by a clause terminal F0 rise, followed by a pause. Using this rule, it is possible to disambiguate sentences that would be ambiguous otherwise, but the boundary signalling is reported to be overly marked and somewhat exaggerated. The same paper also explored means of signalling coherence. Strong coherence signalling between two words may make the

[1] It should, however, be noted that no F0 *values* are implied in the system.

listener perceive a boundary between two others. Merging of phonemes, reducing duration and lowering of F0 level – in effect realizing a deaccentuation – and introducing a focal accent were all successful means of disambiguating the sentences. This indicated that coherence signalling might be as important for phrasing as is explicit boundary signalling. In Bruce and Granström (1989), another way of realizing sentence internal clause boundaries was introduced. Based on recordings of a male Stockholm Swedish informant, a new model for the tonal realization of phrasing was developed. Sentence internal boundary signalling was now modelled with a decrease in F0 before the boundary, a moderate reset after the boundary and differences in the prominence relations of the words related to the boundary.

5.3.3 Focal accent and autosegmental phonology

Bruce and Granström (1989) also added another feature to the model: a timing difference between focal accent in compounds and non-compounds. In compounds the focal rise is critically timed to the secondary stressed syllable, whereas in non-compounds the rise appears to be less constrained. This was based on earlier work by Bruce (1987), where it was shown that the only critically timed part of a focal accent in Swedish is the beginning of the stress group. The work up to this point is summarized in Bruce and Granström (1993), where these issues of prominence and phrasing are unified in a single model. A model of duration developed in Carlson and Granström (1986, 1989) was also integrated. At this stage, the model consisted of a number of pitch accent elements with turning points (maxima and minima), synchronization points and associated prominence levels. The model was now formally formulated in accordance with autosegmental phonology (a non-linear type of phonology) and is presented in Table 5.1.

Table 5.1: Features of the intonation model in Bruce and Granström (1993).

Category	Turning points
Unaccented	–
Accent 1	HL*
Accent 2	H*L
Focal Accent 1	HL*H
Focal Accent 2	H*LH
Focal Accent 2 (compound)	H*L. . . L*H
Initial juncture	L
Terminal juncture	L

Non-linear phonology differs from the classical linear phonology which states that the phonemes in a language are building blocks that occur sequentially, never

on top of each other or in an overlapping fashion. In contrast, in non-linear phonology there are more than one sequence with phonological elements.

Within non-linear phonology there are two main branches; the *metrical*, where stress and speech rhythm is treated (see Chapter 3), and the *autosegmental*, which deals with tones and accents. The particular feature of the latter is that it introduces a tonal layer, which is separate from the segmental-phonemic one. There is thus one sequence with vowel and consonant phonemes and one sequence with tonal information. There are also directions on how the tones may link to the phonemes.

In these models, the H and L represent a High and a Low tonal turning point, respectively. The * notation signals that there is an association between a tonal and a segmental (phonemic or syllabic) element. In this way the positions of the turning points are specified in relation to the elements in the segmental layer. It is, however, not necessary for all tonal elements to be associated with specific positions in the segmental layer. This enables "floating" tonal elements whose positions are determined only by means of their relationship with other elements in the tonal layer. They are floating since no particular segment can be identified as the carrier of the tone. The focal accent of Standard[2] Swedish is an example of an such an element.

5.3.4 Discourse and dialogue

Another addition to the intonation modelling of Swedish was initiated with the PROZODIAG project (Bruce et al., 1995, 1997, 2000). So far, the model had mainly been developed using single-utterance laboratory speech. The intonation model was now extended to cover dialogues and multispeaker conversations. This involved incorporating information on lexical semantics on the discourse level. Several global components, involving parameters for F0 register and F0 range, were added to the model. By changing the values of these parameters from phrase to phrase, F0 contours that were more similar to those found in dialogue speech could be generated.

An important methodological step in the research program was the development of model-based resynthesis. The model enabled F0 contours to be calculated on the basis of symbolic utterance-level representations with pitch accents and boundary tones. Differences between the model-generated and actual original F0 contours were then used to fine-tune the model.

[2]Note that the term 'Standard' is somewhat controversial, since it implies that there is something standard about one specific dialect and that all other dialects are variations, even degradations, of it. We mean the dialect that in the terminology of Bruce and Gårding (1978) is called 'Svea' or 'East', and which is spoken in the region in and around Stockholm. We still use the term 'Standard' for this dialect since it is a rather accepted term for this dialect, at least in the context of Swedish intonation modelling.

5.3.5 Model-based resynthesis

In Bruce et al. (1995), the model was extended with more specific temporal and F0 level information. This was partly a result of the growing general interest in analysis-by-resynthesis. With this technique, new developments of intonation models may be tested almost instantaneously by immediately getting audible feedback on the applicability of a new rule through resynthesis of an existing utterance. This, in turn, originated from recent developments in speech signal processing, e.g., the PSOLA technique (Moulines and Charpentier, 1990), which now allowed prosodic manipulations of pre-recorded speech to be performed very fast and with very high signal quality of the resulting speech. With earlier LPC-based methods it was also possible to alter the intonation, but not without a significant decrease in quality of the speech signal. This development increased the need for explicitness of the model, as computational techniques became the major means of testing the model. See Table 5.2.

An additional set of remapping rules was also developed in order to capture some of the variations of the general patterns that occur in production data. One such rule realizes the L* of a focal Accent 1 as upstepped, producing the sequence 'H H* HF'. Another rule concerns the occurrence of plateaus in post-focal position: the first post-focal H becomes an HF. The pitch level thus remains at the same level as the focal gesture.

5.4 Lyberg (1981)

A computational scheme for the generation of F0 points from syllabic input was suggested by Lyberg (1981). His system produces F0 contours for declarative sentences of Standard Swedish by assigning F0 maxima and minima on accented words. The system uses information about the location of syllables with main stress in accented words. Accent 2 words get one maximum point and one minimum point, whereas Accent 1 words only get one minimum point. There are also special rules for phrase-final syllables. By ordering the rules in a manner similar to a tree structure, F0 points may be generated by examining the properties of each syllable and collecting the outputs of the rules. Full contours may then be produced by connecting the points through the use of, e.g. a cosine function.

The features which distinguish this model from the Carlson and Granström (1973) model are the separate rules for phrase-final syllables, that a syllable's position in the phrase has an effect upon the placement of points, and the apparent lack of a declination line. The lack of a specific declination component makes the model somewhat similar to the early model by Bruce (1977), in that the F0 contour is modelled by a sequence of tonal points.

Table 5.2: Summary of the rules for splitting transcription lables into turning points, and for distributing the turning points temporally. HF is used to indicate the level of the target for a focal gesture.

1. Accent 1, **HL***
   H 100 ms before the label
   L* at the label

2. Accent 2, **H*L**
   H* 30 ms after the label
   L one half of the distance in time to the next label

3. Focal Accent 1, **(H)L*H**
   H 100 ms before the label
   L* at the label
   HF two thirds of the distance in time to the next label

4. Focal Accent 2, **H*LH**
   H* 30 ms after the label
   L one third of the distance in time to the next label
   HF two thirds of the distance in time to the next label

5. Last part of focal Accent 2 compound, **L*H**
   L* at the label
   HF one half of the distance in time to the next label

6. Phrase final, **LH%**
   L 100 ms before the label
   H% at the label

7. All others, **%L, %H, L%, H%**
   At the label

5.5 Linguistic preprocessing for intonation

An attempt at integrating a linguistic preprocessor with a text-to-speech system was developed by Horne and Filipsson (1996). They modelled the relation between the information status of a word and its degree of prominence. A referent tracker that marked all content words as either "new" or "given" was developed. This tracker used means of coreference through lexical-semantic relations in order to find referents that had been mentioned earlier in the current discourse. They then developed a set of rules to assign focal accent on words representing "new"

information, but not on those labelled as bearing "given" information. The tracking of these lexical relations was implemented for predicting accent assignment within a restricted domain.

Horne and Filipsson (1994) and Lindström et al. (1995) developed a text analysis component for the generation of prosodic structure for Swedish texts. They used syntactic analysis, prosodic parsing algorithms and a part-of-speech tagger in order to produce a prosodic hierarchy consisting of prosodic words, phrases and utterances.

5.6 INTRA

Related work on intonation modelling is described in Frid (1999), where an interactive speech transcription tool (INTRA) and a system for the general description of intonation rules (Intonation Description Language, IDL) are developed. With this tool, utterances may be segmentally and tonally transcribed, and the transcriber may get instant feedback on the accuracy and adequacy of the transcription of the utterance by synthesizing a speech waveform on the fly with the segmental and tonal transcriptions as input. This synthesized speech sound may then be examined auditorily. Transcription labels may be moved by simple drag-and-drop, and the tonal transcription is related to the actual displayed pitch contour through a model of F0 contour generation, which is immediately updated as labels are added, deleted or changed. Different phonetic realizations of the tonal labels are possible, a fact which has been utilized to implement different dialectal variants of Swedish intonation.

5.6.1 Intonation modelling

The intonation component of INTRA is intended to be language-independent. Therefore, we have developed an intonation description formalism, inspired by the intonation model presented in Bruce et al. (1995). This formalism includes the following components:

- User-level (phonological) labels, which represent the linguistic features of the language or dialect. By convention, the accent labels are positioned at the vowel onset in stressed syllables. Boundary labels are placed at the end of the phrase.
- Rules for converting user-level labels into tonal turning points (TTPs).
- The temporal specification of a TTP is specified as an absolute value in milliseconds relative to the underlying user-level label or as a relative distance in percent depending on the position of the next TTP label.

- Specification of the target F0 levels for all the TTP categories.
- Context-dependent remapping rules may change, delete or add a TTP in order to take care of the spreading of tones and downstepping.
- Global parameters: phrase start and end F0 levels.

Each language or dialect is implemented by writing rules that specify the above components and each model specification is stored separately. In (1) is an example of the rules for producing the word accents of Standard Swedish. These are both realized as an F0 fall, but they have different timings relative to the onset of the vowel in a stressed syllable.

(1) HL* {H -100} {L 0}
H*L {H 30} {L R 50}

The HL* (Accent 1, acute) is realized by reaching a H level 100 ms before the vowel onset, and a L level at the vowel onset. The H*L (Accent 2, grave), on the other hand, is realized by reaching the H level 30 ms after the vowel onset, and a L, which comes at one half of the distance to the next label (the 'R 50' meaning "50% of the distance to the next label").

The actual F0 levels are then specified by stating explicitly the target level of each TTP category, as in (2).

(2) L 110
H 150

Since all rule sets are independent and the rules are reconsulted at every change in a user-level label, intonation models are easily switched. By using multiple tiers with tonal labels and assigning a different intonation model to each tier, it is even possible to use – view the results of, and synthesize with – different models simultaneously.

Downstep

Downstep may be initiated by the rule:

DOWNSTEP A B1 N1:B2 N2

This statement tells the system that when encountering the label A, it should downstep all occurrences of the label B1 with a factor of N1. A and B1 are TTPs, whereas N1 is a number between 0 and 100, that gets converted into a fraction by dividing it by 100. There may be more than one label specified for downstep. This is then expressed with a list of labels and downstep factors, separated by colons, as follows:

DOWNSTEP A B1 N1:B2 N2

In order to prevent downstep across phrase boundaries the following rule may be used:

DOWNSTEP_BREAK A

This means that the process of downstepping should be discontinued when encountering the label A. In practice, this is often 'L%'.

Remapping rules

The remapping rules are implemented by rules of the form:

REMAP A [B] C = D

The meaning of these rules is 'change the item between [] to the item after the = if the item between [] appears in the context of the items to the left and right of the []'. A, C and D may be empty, but B must contain a symbol. A, B and C are user-level labels, whereas D is a TTP.

5.6.2 Dialects

The intonation model was developed so that it would be possible to generate the different intonational characteristics of various dialects, without changing the phonological (tonal) labels. This is done by making the mapping from phonological transcriptions into F0 events dialect dependent, as described in the previous section.

There has been a recent increase in dialect research in Sweden triggered by the SWEDIA 2000 project (Bruce et al., 1998), and some of the major dialects of Swedish are included in the system as models for the F0 generation. Thus, we may simulate the intonation of different dialects. The result is a good approximation of how close we can get to producing dialectal variation by varying intonation only and it illustrates the role that intonation plays in differentiating different dialects of Swedish.

However, segmental differences, such as diphthongs and the realization of the /r/ phoneme, are currently not dialect specific as the diphone database utilized does not contain all the desired dialectal variants. A dialect-independent database is a possible future extension.

For Swedish, we have followed the dialect typology and realization rules in Bruce and Gårding (1978), which have been further elaborated with temporal specifications. This typology is based on prosodic characteristics and identifies five different main dialect categories of Swedish: Svea (EAST), Göta (WEST),

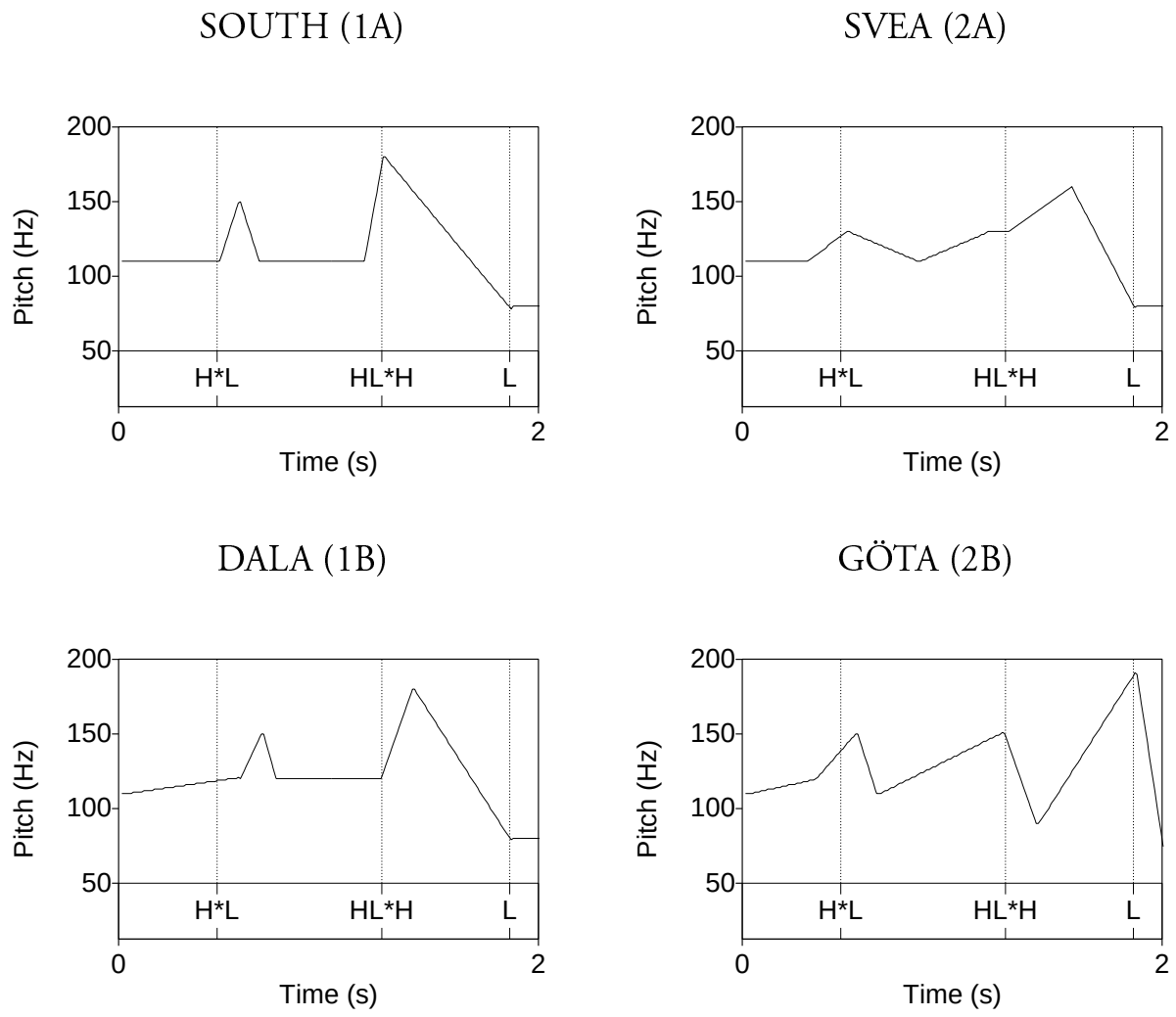

Figure 5.1: Pitch contours of four different dialects generated using the same phonological labels. Different acoustic realizations are obtained by using different rules for each dialect.

Southern (SOUTH), Dala (CENTRAL) and Finland Swedish (FAR EAST). The first four dialects were then interpreted in terms of the intonation model. The model identifies a number of discrete categories with associated labels. The model recognizes two levels of prominence, and for each level of prominence the distinction between the two word accents in Swedish. It also includes the accent pattern of compounds, as well as terminal juncture (boundary) tones.

Figure 5.1 shows the different realizations of the four dialects Svea, Göta, South and Dala. The labels 2A, 2B, 1A and 1B are the same labels as used in Bruce and Gårding (1978). In all the dialects, the temporal distinction between the word accents is maintained, but the accent fall is timed differently relative to the vowel onset in the stressed syllable. The Svea dialect has the earliest timing, followed by Göta, South and Dala. Note that the labels 'H*L' and 'HL*H' is somewhat misrepresentative of the F0 contour for the 1A, 1B and 2B dialects. They should only be interpreted as 'Accent 2' and 'Focal Accent 1', without any implications as

to how these tones are realized.

The focal accent label HL*H is realized by the rules presented in Table 5.3, and their different phonetic realizations can be seen in Figure 5.1.

Table 5.3: TTP realizations of the accent label HL*H for different dialects of Swedish. The labels H, HF, LF and L refer to different pitch levels.

Dialect	Realization as Tonal Turning Points
Svea	{H -100} {H 0} {HF R 50}
Göta	{H -20} {LF R 25}
	REMAP LF [L%] = HF
South	{L -100} {HF 0} {LF 100}
Dala	{L -20} {HF 140} {L 230}

Recall that the numbers following the TTPs denote the timing relative to the position of the phonological label. For Svea and Göta, focus is thus realized by having an extra high (HF) after the word accent. Svea has an earlier timing of this high, whereas in Göta the high should come phrase-finally, hence the L% is remapped as a HF if it follows an LF. For South and Dala, focus is realized by having higher highs (HF) and lower lows (LF, South only).

The levels of the TTPs are then specified similarly for all dialects, as shown in (3).

(3) L 110
H 150
HF 180
LF 90

In this way, there are different realizations for each prosodic category in each dialect. These rules have been tested previously by means of resynthesis from hand-made pitch contour stylizations (Bruce and Gårding, 1978), but the implementation in a generative fashion as presented here is novel. By combining the intonation model component with the synthesis methods described above, it is now possible to test the rules on arbitrary utterances.

5.6.3 Discussion

The IDL system may be seen as an instance of an abstract prosodic mark-up language, a model of the prosody generation process itself. In recent works by Shih et al. (2001); Kochanski and Shih (2003), another such system, the Soft TEMplate Mark-up Language (Stem-ML) is described. This notational device was designed as a meta-language, where one of the aims is to be able to accommodate several different intonation theories. It defines a set of tags that may be used for

describing prosodic events such as phrase curves, accents, properties of accents and the interaction between different components. Each tag also has a mathematically defined algorithmic part that generates F0 contours based on the parameter settings of the tags.

5.7 Superpositional modelling of Swedish intonation

Gårding (1983) analyses Swedish intonation with a model where lexical prosody is separate from phrase or sentence prosody. A *tonal grid* serves as the reference for local F0 movements. The shape of the grid is determined partly by sentence mode and partly by the occurrence of *pivots* at major phrase boundaries. Phrase and word accents are then realized by the insertion of highs and lows on the grid. The position and height of the tones are thus partly determined by the non-local tonal grid.

Swedish has also been analysed using Fujisaki's command-response model (Fujisaki and Sudo, 1971; Fujisaki et al., 1993; Ljungqvist and Fujisaki, 1993, see also Chapter 6). They analyse Swedish intonation in terms of a global, or *phrase*, component, with an initial rise and then a gradual fall, and one or more local, *accent*, components that occur on accented syllables. The phrase and accent commands are modelled by the parameters of the model. These include the number of phrase and accent commands, and the amplitudes and timings of each different command. By using an analysis-synthesis approach, where the best approximation of an observed F0 contour is derived, values for the parameters of the model that are appropriate for Swedish are obtained. The model is reported as being able to successfully reproduce the patterns of Accent 1 and Accent 2 words of Swedish, as well as focus accent in a longer sentence. Prototypical patterns for the different accents are then conjectured.

They also sketch the possible implementation of the model in a text-to-speech system, where specification of word accents, stressed syllables, sentence focus and phrase boundaries may be used to generate the desired model parameters.

Another attempt at superpositional modelling has been undertaken by Fant and Kruckenberg (2001b,a); Fant et al. (2002). Their model addresses both analysis and synthesis of F0 contours across speakers, and uses normalization methods in the frequency as well as the temporal domain. The model assumes an underlying base contour that is modified by accentual contours.

In order to represent word accents, the notation system of Bruce (1977) and Bruce and Granström (1993) is adopted, with some additions. Their transcription system has the notations:

Accent 1	H L* Ha
Accent 2	H* L Hg
Unaccented	Lu

The H and L denote High and Low (like in Bruce's model). The Ha and Hg stand for the focal gesture in acute and grave words, respectively, whereas Lu is used for unaccented syllables. The word *margarinlåda* 'box of margarine' would be transcribed as:

mar	ga	rin		lå	da
Lu	Lu	H*	L	Hg	Lu

Analysis is performed by sampling (measuring) the F0 level at specified positions. These positions are derived from the phonological tonal transcription of the words in the speech fragment under analysis. Main stressed syllables of accented words are sampled at two points, whereas all other syllables are sampled only once. In the example above, there would be six measurement points (one for each syllable + one extra for the accented syllable). In effect, this amounts to a kind of temporal normalization, as each syllable gets represented by one or two measurement points.

All measurements are normalized based on the speaker's average F0 in unstressed syllables (the Lu sample points), and a set of base contours is derived from these Lu points, currently one for clause initial phrases, and one for non-initial phrases. When the model is used for the generation of F0 contours, local modulations of the base contours are overlayed at the accented positions (the H, L*, H*, L, Ha and Hg points). The heights of these modulations are determined partly by the observed heights of these points, and partly by the value of a continuously scaled prominence parameter (Rs) that is derived from features such as word class and position in the phrase, i.e. a linguistic preprocessor. Finally, a declination function is realized by scaling the modulation height according to the temporal position in the phrase.

The authors report that synthesis experiments have been promising and that the model is able to produce F0 contours that are very similar to naturally occurring ones.

5.8 Discussion

Several different intonation models have been suggested for Swedish. Perhaps the complex nature of the interplay between word accent and focal accent has contributed to this. According to Ladd (1996), the relation between global and local intonation events has been described using two types of models: *linearity* and *overlay* models. Linearity models describe intonation as a sequence of categorically distinct elements that are chosen from a standard set of discrete tonal movements.

Each tonal element then has a certain phonetic realization and by concatenating these the appearance of the F0 contour is obtained. These models are also called tone sequence models, as they consist of sequences of tonal elements. In these models there is no real difference between global and local intonation patterns, it is rather a question of the length of the domain that a certain tonal element operates upon.

In overlay models, the pitch contour is treated as a complex function, which may be split up into simpler functions. There is a global component, which is effective over a long domain, and local components, that are superimposed on the global component in order to represent more rapid pitch changes. The tonal elements thus interact in order to produce the final F0 contour. Superpositional models may have a weaker grounding in phonology, and instead use speech physiology as a starting point. But they should not be excluded from the autosegmental point of view. The phrase and accent commands are, after all, tonal elements on different levels.

Chapter 6

Intonation: acoustic models & stylization

6.1 Introduction

In this chapter, we review some of the leading theories of acoustic intonation modelling. Special attention is given to the field of stylization of real intonation contours. Intonation models come in many different variants. There are physiologically oriented models, phonological models and perceptually inspired models. Another type is represented by acoustically oriented models that attempt to describe and reproduce the patterns we see in the intonation analysis of natural speech through stylization. Stylization uses intonation from spoken utterances, and extracts information from them. Extracting this information is often done by reducing the intonation curves into their most significant components.

After introducing some of the major models, we perform a small comparison of two of the stylization models. Finally, we discuss the applicability of stylization models for intonation experiments and motivate the selection of two of the models for our further work in this thesis.

6.2 IPO

One of the oldest and most widely known methods of intonation modelling is the IPO model (Cohen and Hart, 1967; Hart et al., 1990) developed at the Instituut voor Perceptie Onderzoek at Eindhoven, The Netherlands. The background to the model lies in the observation that some F0 movements are perceptually important wheras others are not. The model attempts to capture the relevant melodic changes in speech and to link them to certain functional aspects. It provides a framework for studying physical and linguistic aspects of intonation. The IPO approach may be seen as one of the first successful attempts at combining a phonological

description with phonetic realization and has therefore developed a strong relation with synthesis. Stylization lies at the heart of the model and is a very important aspect.

The main idea is that a model of intonation should be extracted from raw acoustic F0 data by modelling the physical changes reflected in the pitch contour which are perceptually relevant to intonation. In order to find out which physical changes are perceptually important, the IPO method uses a stylization procedure to simplify the original pitch contour without causing any perceptual changes. This stylization procedure consists of an iterative, interactive process whereby a listener replaces the original pitch contour with a minimal number of straight lines but with the overriding criteria that the original and stylized versions must remain perceptually equivalent. In this *stylization* phase, the pitch curve is represented as a sequence of straight F0 segments, each with three measured parameters: semitones of change, milliseconds of duration and alignment with syllable boundaries. The resulting stylized pitch contours are often called "close copy" contours since the stylizations resemble the originals visually and, per definition, are perceptually equal. Next follows a *standardization* phase where the F0 segments are classified into pitch movement features according to their acoustic properties. Each movement is characterized according to the four parameters in Table 6.1.

Table 6.1: Pitch movement features in the IPO model.

Parameter	Possible values
direction	rise/fall
timing	early/late/very late in the syllable
rate of change	fast/slow
size	full/half

After examining a sufficient number of utterances, common features of the contours are collected in order to create an inventory of pitch movements for the language under investigation. An important aspect is whether the movement is *accent lending*, i.e important for perception of prominence, or not. Once the pitch movement inventory for a language is established, a "grammar of intonation" can be defined where the possible and permissible pitch movement combinations in a given domain are described and linked to linguistic functions. For example, in the Dutch grammar of intonation (Hart et al., 1990), intonation patterns are composed of the pitch movements shown in Table 6.2.

When two movements occur on a single syllable, the symbols are combined with an ampersand, e.g., 1&A. The IPO model has been used for synthesis in

Table 6.2: Pitch movements in an IPO grammar for Dutch.

Label	Description
1	early prominence-lending rise
2	very late non-prominence-lending rise
3	late prominence-lending rise
4	slow rise extending over various syllables
5	overshoot after a rise
A	late prominence-lending fall
B	early non-prominence-lending fall
C	very late non-prominence-lending fall
D	slow fall extending over various syllables
E	half fall that may be prominence-lending

(for instance) Dutch (Terken, 1993), English (Willems et al., 1988) and French (Beaugendre, 1994) and also for recognition and automatic labelling of pitch contours (Mertens et al., 1996). Sproat (1998) and Möbius (2001) critise the IPO model on several grounds. The stylization method can yield inconsistent results, since there is no standardized way of creating data. In order to extract the pitch movements of a language many hours of interactive listening and resynthesis are necessary. They also argue that the categorization of pitch movements is not objective, but guided by the goal of obtaining a model that is perceptually assessed, and further that the intonation grammar may produce perceptually unacceptable contours.

6.3 Superpositional intonation models ("Fujisaki-Öhman" models)

The paradigm of physiologically oriented modelling of F0 contours was initiated in Öhman and Lindqvist (1966) and Öhman (1967) with the "Larynx model". The basis for this approach is that all aspects of prosody are contolled by muscle actions. The dominant part of current physiologically oriented intonation modelling is developed in Fujisaki and Nagashima (1969); Fujisaki and Sudo (1971); Fujisaki and Hirose (1983). F0 contours are described as the result of superposition of *phrase* and *accent* components. The components are generated by second-order, critically damped linear filters in response to phrase commands and accent commands. The phrase command is an impulse function, whereas the accent command is a step (rectangular) function. The combination of commands and filter outputs has

resulted in the model often being called a *command-response* model. It is grounded in physical modelling of the vibrations induced in the vocal tract during speech. By decomposing the F0 contour of an utterance into the components of the model a parametric representation of the utterance is obtained. This representation consists of the timings, amplitudes and damping factors of the commands and this may also be used for reconstruction, or, by using a different set of values, for synthesis of F0 contours. The model has been applied to several languages, including Japanese (Fujisaki and Hirose, 1984), English (Fujisaki and Ohno, 1995) and Swedish (see Chapter 5).

According to Taylor (2000), the Fujisaki model and the Tilt model (see Section 6.6) have many things in common, most notably that they both model accents as events without any categorical accent types inherent in the model and that they are both formal enough to be suitable for speech synthesis. One of the problems of the Fujisaki model is the requirement, due to the phrase component, that the model always contains a rising part followed by a falling part. Another is that accents always have the same shape, differing only in duration and amplitude, whereas natural speech shows more variation in the appearance of pitch accents.

6.4 ToBI

ToBI (Tones and Break Indices) (Silverman et al., 1992) is a system for transcribing intonation and prosodic structure, originally designed for American English utterances, but later applied in other languages and dialects as well (see below). It may also be seen as an intonation model based on the Autosegmental (Goldsmith, 1976) and Metrical theories of intonation (Liberman, 1975; Liberman and Prince, 1977), developed by Pierrehumbert (1980) and Pierrehumbert and Beckman (1988) (see also Ladd, 1996). It is not strictly a stylization model even though one of its aims is to provide an acoustic description of intonation. It still deserves some attention in the present context, since its influence on intonation research and synthesis has been very large.

The development of ToBI was, at least partly, triggered by an increasing need to label prosody in a standardized way. The tonal inventory was developed in Pierrehumbert (1980) and is used to describe three different intonation events: pitch accent, phrase accent and boundary tone. The model describes tonal categories using two pitch levels, High (H) and Low (L). These are linked to stressed syllables and phrase boundaries. Accents can be monotonal or bitonal,and diacritics are used to indicate the type of event. A '*' is used for pitch accents, '-' for phrase accents and '%' for boundary tones. The full inventory of ToBI is given in Table 6.3.

In a combined tone, e.g., a bitonal accent, the '*' indicates the alignment with the syllable. In addition to this, '!' is used to indicate *downstep*, a lowering of the pitch level of an event in a specific context, and a modification of the acoustic

Table 6.3: The ToBI inventory.

Pitch accents	
H*	peak accent
L*	low accent
L*+H	scooped accent
L+H*	rising peak accent
H+!H*	downstepped accent
Boundary tones	
L%	low final boundary tone
H%	high final boundary tone
%H	high initial boundary tone
Phrase accents	
L-	low phrase accent
H-	high phrase accent

interpretation of a label. The individual syllables of an utterance are then tagged using this inventory, and the resulting transcription enables both identification of important events and a rough description of the shape of the F0 contour. The transcription system is thus ambitious in that it tries simultaneously to describe the significance of intonation events and to capture the actual acoustic realization of the events, in terms of the pitch contour. Prosodic coherence between words is annotated with numbers ranging from 0 to 4 (Break Indices), which should be interpreted such that the higher the number, the stronger the boundary.

ToBi has been widely used both in synthesis (Ross, 1994; Black and Hunt, 1996) and analysis (Ostendorf and Ross, 1997). It has also been used in the analysis of other dialects of English (Mayo et al., 1997) and a number of other languages, e.g. Japanese (Campbell and Venditti, 1995), Dutch (to appear in Gussenhoven, 2003) and German (to appear in Grice et al., 2003).

In recent years, the use of ToBi in speech technology has received criticism for introducing a *quantization error* and for lacking granularity (Batliner et al., 2001). Furthermore, Wightman (2002) points out that very few studies use ToBi *as is*, but rather with modifications that make the system language dependent or reduce the symbol inventory; also the descriptive nature of the labels seems to have little to do with a listener's perceptual experience of an utterance.

6.5 INTSINT and MOMEL

INTSINT is a coding system of intonation developed by Daniel Hirst and colleagues (Hirst et al., 1991). It has been tested on many languages, e.g., French, Spanish, English, Arabic and Swedish (Strangert and Aasa, 1996). The symbolic coding is expressed in terms of tonal target points, which are defined according to the speaker's pitch range and the relative height of the preceding and following target points. There are three absolute tones, which are listed in Table 6.4, and three relative tones, which are listed in Table 6.5.

Table 6.4: Absolute tones in the INTSINT model.

T	Top of the speaker's pitch range
M	Mid of the speaker's pitch range (also used in initial position)
B	Bottom of the speaker's pitch range

Among the relative tones, a distinction is made between iterative and non-iterative tones, since iterative tones appear to use a smaller F0 interval than non-iterative ones.

Table 6.5: Relative tones in the INTSINT model.

Non-iterative	
H	Higher Target is higher than its immediate neighbours
L	Lower Target is lower than its immediate neighbours
S	Same Target is not different from preceding target
Iterative	
U	Upstep Target in a rising sequence
D	Downstep Target in a falling sequence

In Hirst et al. (1998), the authors emphasize that it is not necessary to know the inventory of pitch patterns in a given language in order to use this system, and that the system is rather like IPA, providing a means of gathering data on the basis of which phonological descriptions may be further specified.

The actual stylization, which can also be seen as an intermediate representation between the coding system and the pitch curve, is done by an automatic procedure called MOMEL (Hirst and Espesser, 1993). The method works by selecting target turning points in the F0 contour through an elaborate procedure and then connecting them using a quadratic spline function, which is a completely smooth function. In this way, "angles", which occur when connecting targets using straight lines are avoided. The model has been used in synthesis of for instance, French and Italian (Campione et al., 2000).

The necessity of avoiding using straight lines has been questioned by Hart (1991) since results from perception experiments showed that subjects cannot hear the difference anyway. Hirst et al. (1998) defend the model on the grounds that it still produces curves which are more similar to original F0 curves and that it can provide a more economic representation. However, Strangert and Aasa (1996) reported that the model sometimes produces a too heavy smoothing of the intonation contour.

6.6 The Tilt intonation model

A recent model, which has quickly become very influential, is the Tilt theory of intonation. It has been developed by Paul Taylor (Taylor, 2000) and it aims at a representation of intonation oriented both towards speech synthesis and intonation analysis. It is a phonetic model of intonation that uses continuous parameters to characterize intonation events. Intonation is represented as a linear sequence of such events, which can be pitch accents or boundary tones. Each event is described by its temporal position, the fundamental frequency at the start of the event, duration, F0 amplitude and the "tilt" parameter. The tilt parameter is a dimensionless parameter that expresses the overall shape of the event, based on the relationship between the rising part and the falling part of the event, disregarding amplitude and duration.

The model can be used both for synthesis and analysis by defining a reversible function linking the F0 curve through a sequence of intonation events to the syllabic nucleus and the acoustic parameters. An illustration of the model is shown in 6.1.

In recognition and automatic pitch analysis, the tilt model is often used together with an *intonational event detector*, also described in Taylor (2000). The tilt model has no way of detecting these events on its own, but provided a sequence of events, it can compare the event sequence with an F0 contour and characterize it in terms of events using the acoustic parameters described above. In Dusterhoff (2000) the performance of the model is evaluated by generating F0 contours from descriptions obtained from tilt analysis and comparing them with the original contours. For two large American English databases, the mean correlation was above 93% and the Root Mean Square Error (RMSE) less than one third of the standard deviation

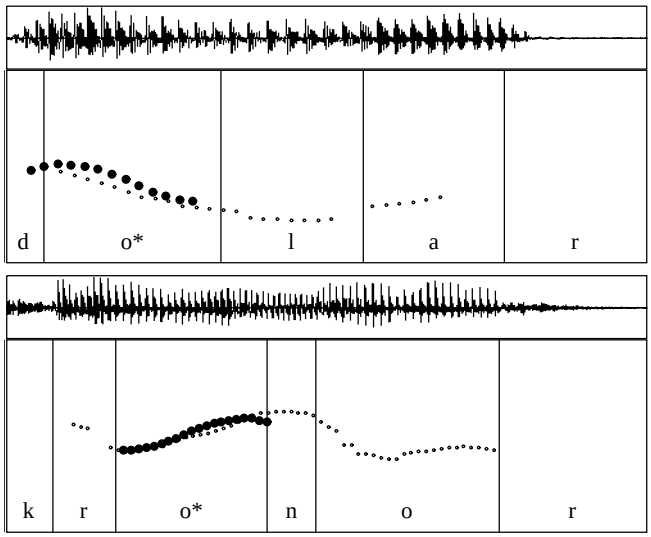

Figure 6.1: Illustration of the Tilt model. Displays show (top) the the A1 word 'dollar', and (bottom) the A2 word 'kronor'. Actual pitch contours are shown with small dots, tilt data at the right of the diagram, and resynthesised contours from the tilt data with large dots. Note that the search parameters has been limited to the vowel boundaries, and therefore the F0 contour is only reconstructed between these boundaries.

of the natural F0 variaton in the databases.

For synthesis, experiments are described in for instance Dusterhoff et al. (1999) and Dusterhoff (2000). Decision trees (see Chapter 4) are used to predict tilt parameter values for syllables from suprasegmental variables such as lexical stress, position within a phrase and relative positions of tilt events. Comparisons between original contours and reconstructed ones show lower variations than the natural F0 variation of the tested speakers. They conclude that the model produces reasonable intonation and that it is possible to achieve a high correlation between original and synthetic F0.

A problem with the tilt model is the inability of events to code complex tones, like the focal accents in Standard Swedish, which consist of a fall followed by a rise. Such a pitch accent would have to be coded using multiple events. Furthermore, the model might be accused of being just F0 data reduction and lacking linguistic interpretability. However, Taylor (2000) claims that for example F0 amplitude may be related to perceived prominence.

6.7 "Contour faithful" perceptual stylization models

In Section 6.2 we described the IPO model, which is an attempt to stylize F0 contours by approximating them perceptually. A slightly different approach is

taken by models that, in one way or another, aim at having a representation that is as faithful as possible to actual F0 contours, and at the same time are inspired by theories of tonal perception. They attempt automatic stylization based on perceptual criteria.

6.7.1 Stylization based on tonal perception

The history behind stylization comprises several different attempts. One of the most significant is D'Alessandro and Mertens (1995), who describe a perceptually grounded model of intonation stylization. They use perceptual criteria to localize the points of an intonation curve which are most important for the perceptual impression of intonation in a sentence. A fundamental idea in the model is the inability of the auditory system of humans to follow too rapid changes in the fundamental frequency in speech. The authors base this view on experiments with perception of vibrato in singing, where individual pitch changes were found to be unimportant and instead an integrated model (WTAM) could represent the perceived pitch better. Moving on to perception of intonation in speech, they introduce two important psycho-acoustic concepts for the audibility of pitch changes: the *glissando threshold* and the *differential threshold* of pitch change (cf. the works by Rossi (1971, 1978) on glissando and Hart (1976) on proximity of changing tones).

The glissando threshold determines how large a change in frequency must be in order to be perceived as a change. The glissando threshold is also dependent on duration, and basically it determines whether a change in frequency over time is large enough to be a slope or not. The rate of the frequency change over time is called the *glissando rate*. It is measured in *semitones per second*. As the semitone scale is logarithmic, this makes it independent of the absolute frequency. The differential threshold is the minimum difference in slope that is necessary to distinguish between two successive glissandi. Using these two auditory parameters, the stylization model for speech prosody is then developed.

Another important observation is that changes in the spectral properties of the signal tend to function as boundaries (House, 1990), breaking up a voiced continuum into a sequence of syllabic nuclei. The speech signal is therefore first segmented into syllable-sized segments based on spectral change and energy variations. A further decomposition into tonal segments is then performed based on the two thresholds. This segmentation is done in two steps: first a recursive step where segments are split, then a sequential step where segments may be joined again.

1. Find the largest local difference between:
    - a straight line connecting the start and end points of the current segment

- the observed pitch curve

2. If the difference is larger than 1 (one) semitone and the glissando rate is larger than the glissando threshold then split the segment and recur (repeat from 1 for each subsegment).

This procedure results in a sequence of segments, all with a uniform slope. The second step now examines the segments from left to right. If the difference in slope between two contiguous segments is below the differential threshold the segments are joined again. Finally, target values are assigned for each tonal segment by selecting the pitch values at the boundaries of the tonal segments.

The tonal segment of a syllable-sized segment may thus be static or dynamic, based on whether the pitch change in the segment is above the glissando threshold or not. A dynamic tonal segment may then be a simple or a compound tonal segment. A simple dynamic segment is one tonal segment with a perceived pitch change. A compound dynamic segment consists of two or more tonal segments and at least one has a perceivable pitch change.

The model has also been tested in a perception experiment by the authors. This was done by comparing original F0 contours with resynthesized versions. The resynthesis was done from stylizations which were obtained by analyses according to the model. 20 listeners judged a total number of 3776 stimuli (about 188 each). Stimuli were produced by using different values of the auditory thresholds. The task of the listeners was to determine for a pair of stimuli whether they were the same sentence or not. For the best set of threshold values, they got 78.92% judgements as "same", meaning that in around four out of five cases the original and resynthesized versions were judged perceptually identical. The conclusion is that, despite a few errors, automatic stylization with a high level of perceptual equality is possible.

6.7.2 Point removal stylization

Another kind of stylization procedure is provided in the speech analysis program PRAAT (Boersma and Weenink, 2003). The following description of the method is provided by Paul Boersma (p.c.). Pitch points are collected in a data structure called the PitchTier. The stylization function removes points from this data structure. The two points at the edges are never removed. The point that is closest to a straight line connecting its neighbours is removed. This is repeated until all points lie within a certain distance (e.g. 2 semitones) from the straight line between their neighbours. Let's say we start with the following sequence of pitch points, all at the same equal time distance:

100 - 95 - 80 - 105 - 90 - 110 - 125 - 125 - 100

The point that will be removed first is "110", which is at a distance of 2.5 Hz from the line between its neighbours 90 and 125 Hz. This results in:

100 - 95 - 80 - 105 - 90 - – - 125 - 125 - 100

Then we remove the "95" (distance 5 Hz) and we get the following list:

100 - – - 80 - 105 - 90 - – - 125 - 125 - 100

Then the second "125" (distance 12.5 Hz):

100 - – - 80 - 105 - 90 - – - 125 - – - 100

The 80 now lies at a distance of 23.3333 Hz from the straight line between 100 and 105, the 105 lies at a distance of 20 Hz from the line between 80 and 90, and the 125 lies at a distance of 30 Hz. The process is now finished, since all these distances are more than 2 semitones. This method ensures that the entire PitchTier is within 2 semitones from the original.

In one sense, this method follows the IPO Hart et al. (1990) paradigm of intonation stylization in that it tries to approximate F0 contours using "close-copy stylizations". This method is, however, fully automatic, and does not use the interactive step.

6.7.3 Other methods of stylization

Scheffers (1988) developed a stylization method where candidates for stylization points are calculated using linear regression between a set of successive F0 values. By comparing predicted and actual F0 values a set of stylization points is obtained. A listening experiment showed that stylized contours with variations of up to 1.5 semitones were indistinguishable from the original contours.

House (1990) used a model where the speech signal was segmented into tonal segments using an intensity-based method. Average F0 values during the beginning and end of each segment were then calculated. An interpolation between these two values resulted in a linear stylization. Listening tests showed that original and stylized contours were often indistinguishable, with the possible exception of phrase-final rise-fall configurations.

6.7.4 Comparison

The D'Alessandro and Mertens (1995) and the PRAAT (Boersma and Weenink, 2003) methods are apparently very similar. The first one approaches the F0 contour by starting with just a straight line, and then adding pitch points from the original

contour at the temporal locations where the original and the stylization differ most until this difference becomes lower than a given threshold. The second method instead starts with the original pitch point curve and removes points where such a removal causes the least change in the pitch curve. This is repeated until the change becomes too large. Despite approaching stylization from two apparently different directions, these methods return very similar results. We made a comparison of the method described by D'Alessandro and Mertens (1995) and the PRAAT method. We used 485 sentences taken from the Ekot corpus (read speech, Swedish radio news, see also 7), performed a pitch analysis and then tried the two different methods on each sentence. Numerical comparisons are presented in Table 6.6.

Table 6.6: Comparison between D'Alessandro and Mertens (1995) and PRAAT (Boersma and Weenink, 2003) stylization methods.

	Mean	Min	Max
RMSE (Hz)	1.17	0.22	2.84
Correlation	0.9978	0.9859	1

The table shows the average difference for all utterances, as well as the minimum an maximum RMSE values for any utterance. The maximum RMSE for a sentence between the two models is 2.84 Hz and the correlation is always very close to 1. This shows that the two stylization methods yield almost identical results. In fact, the difference between two different stylizations of the same sentence is often smaller than the difference between either of the models and the original.

6.7.5 Discussion

In stylization models, any pitch point in an F0 contour is a possible candidate. This can easily be criticized from a perceptual viewpoint. It has been shown (Hart, 1991) that not all portions of the pitch contour are equally important. For example, in regions with a large spectral change listeners are not sensitive to pitch movements, only pitch levels (House, 1990). It is also reasonable to assume that pitch information in stressed syllables is more important than pitch information in unstressed syllables. Such additional segmental information may be used to guide the stylization process so that only points in certain portions of the F0 contour are eligible candidates.

When performing stylization with a perceptual threshold, the number of pitch points remaining after stylization is allowed to vary with the complexity of the pitch curve. This causes problems for stochastic model building, as two words with the same linguistic analysis may have a different number of pitch points, thereby

making it difficult to model them. It is, however, possible to tell the stylizer how many points are required in the end (thereby compromising the perceptual limit). A possible path to take is to perform a perceptual stylization for all words in the same category and calculate the average number of stylization points for this group, and then to request that (rounded off) number of points when stylizing all the words in that group. Another solution is to use the maximum number of points in a group to stylize all words in that group.

6.8 Other models

Other major contributions to intonation modelling include Kohler (1991), van Santen and Möbius (2000), Grønnum (1992) and Möhler and Conkie (1998). For intonation modelling with special reference to Swedish (works by, among others, Gårding, Bruce and Fant), see Chapter 5.

6.9 Synopsis

This chapter has shown that there is a large variety of methods for modelling the acoustics aspects of intonation. As a tentative classification, we would like to suggest that there are three major types:

1. *Acoustic-perceptual* models: the types that describe intonation by being more or less faithful to the original contour. Their descriptions consist of collections of points with timing and frequency information that is actually found in real F0 contours and is extracted using perceptual criteria. The stylization models, INTSINT and IPO fall in this category.

2. *Acoustic-parametric* models: the types that attempt to abstract away from the actual observed contours by parameterizing the shape of the contour. Among these we find the Tilt and Fujisaki models.

3. *Phonological* models: the types that provide symbolic labels that are transparent enough to provide a rough description of the F0 contour and also correspond to pitch movements which replicate it. ToBI is such a model.

One of the purposes of this survey was to provide a basis for the selection of models for the future work in this thesis. In order to select models we apply the following two criteria:

1. the model should not have been tried on Swedish before

2. the model should provide an automatic analysis method

INTSINT has been tested on Swedish previously (Strangert and Aasa, 1996) and the same is true for the Fujisaki model (Ljungqvist and Fujisaki, 1993). Likewise, a ToBI-like transcription system has been developed for Swedish (Bruce and Granström, 1993), see also Chapter 5. The IPO model does not provide an automatic analysis method. In fact, if it did, we think it would be very similar to the stylization models.

Two models remain, one in each of the acoustic categories: the stylization model and the Tilt model. These will consequently be used in the synthesis and recognition experiments in the following chapters. In Chapter 7 we will test the stylization method in synthesis of Swedish intonation and we will then apply both models in the recognition experiments in Chapter 8.

Chapter 7

Prediction of intonation patterns for Swedish content words

7.1 Introduction

The existing model of F0 prediction in our department's systems and tools for speech synthesis (Bruce et al., 1995; Filipsson and Bruce, 1997; Frid, 1999, , see also Chapter 5) is rule-based. It can be described as a 'ToBI-style' intonation model in that it uses tonal turning points, represented by Ls and Hs, which are mapped to time and frequency values that are connected by straight lines in order to produce a pitch contour. The model has been fairly successful at producing a neutral intonation of Standard Swedish (Bruce and Granström, 1993) and has also been applied to different dialects of Swedish (Bruce and Gårding, 1978). There have also been attempts, summarized in Bruce et al. (2000), to incorporate discourse and dialogue features into the model.

Given that recent attempts (Black and Hunt, 1996; Dusterhoff, 2000) at data-driven methods have been rather successful within the area of speech synthesis, and that such approaches, to our knowledge, have not been pursued previously for Swedish, we decided to investigate this technique.

In this study we concentrate on words with an actual realization of word accent. We have not yet tried to predict which words in a phrase get accents, or why some words are deaccentuated. Neither have we accounted for phrasal phenomena like focussing, post-focal downstepping or final lowering. This study is therefore somewhat tentative, rather a test of a possible method for future studies than a full account of Swedish prosody.

7.2 Speech data

The speech data for this study was taken from a corpus consisting of read news (from the hourly Swedish news program 'Ekot'). The speech in the corpus is read at a clear, rather formal style, and the complexity of utterances can be rather high. The mean length of a read sentence is 6.8 s, and the average number of words per sentence is 15.76. The corpus consists of samples from several speakers, both male and female, which all speak a variety of 'Standard' Swedish. The sound quality of the recorded speech is very high and poses few problems for pitch extraction. The corpus currently consists of 300 sentences and from this we extracted about 2300 tokens of content words (nouns, verbs and adjectives) which had a word accent

7.3 Linguistic Analysis

A linguistic analysis was performed in order to classify the words in different categories. Each word's lexical accent was first determined by looking it up in a prosodically annotated pronunciation dictionary (Hedelin et al., 1987). The dictionary also gave information about the number of syllables and the positions of the syllables with primary and secondary stress.

The accent information was then checked manually by listening to the words and determining whether or not the designated labels were correct by using the pitch information of each word. Visual inspection of the F0 was used in some doubtful cases, where the intuitions of the author had to settle the issue.

The accent types used were:

- Accent 1 (acute)
- Accent 2 (grave)
- Compound accent

Note that compounds almost always have Accent 2 (except in some dialects in southern Sweden, see Chapter 8), but a distinction is made here since compounds always have a secondary stressed syllable, which the simplex, non-compound Accent 2 words may or may not have.

In addition to the accent category, the position type of the main stress syllable was also included as a feature used for prediction. The position types were:

- Initial
- Medial
- Final

- Single (for monosyllabic words)

There are many other features that influence the pitch properties of words (Black and Hunt, 1996), in particular the degree of prosodic prominence of a word, but also features at levels above and below the word, such as position in the phrase from the left and right edges, pre- or post-focal position, openness and heaviness of syllables and foot structure. These have not been used in the present study as this would require a much more detailed analysis. Accent type and position type are regarded as important variables that are relatively easy to obtain and are hence suitable for a tentative study like the present one.

7.4 Acoustic Analysis

The goal of the acoustic analysis was to extract pitch information for the selected words in order to build a model for prediction of F0 contours from the linguistic features. Pitch information was extracted from the words by first obtaining F0 contours, then smoothing the contour in order to remove minor perturbations that have little intonational content, and then performing a stylization of the pitch contour.

7.4.1 Stylization

The stylization method was described in more detail in Chapter 6, Section 6.7.2. The stylization works by selecting tonal turning points in the contour. The points are selected so that when reconnecting the points with straight lines, there may not, at any given point along the contour, be a difference in pitch between the reconstructed contour and the original contour that is larger than a set value, in this case one (1) semitone. This results in a series of time/frequency pairs, which describe the contour of a pitch pattern accurately, but with a smaller number of points than the full contour.

The stylization process is illustrated in Fig. 7.1, which shows two pitch contours—original and stylized—of the word 'Helsingfors' spoken by one of the speakers. The orginal contour contains all F0 points obtained from the pitch analysis, whereas the stylized contour has a much smaller number of pitch points.

Note that the points are not anchored to any particular point in the word, neither a temporally aligned one, such as beginning, middle or end of the word, nor a linguistically motivated one, e.g. the vowel onset of the stressed syllable. This might of course introduce some unnecessary variation due to the durations of segments being different from word to word and to the fact that some words begin with voiced segments and others with unvoiced, but we are mainly interested

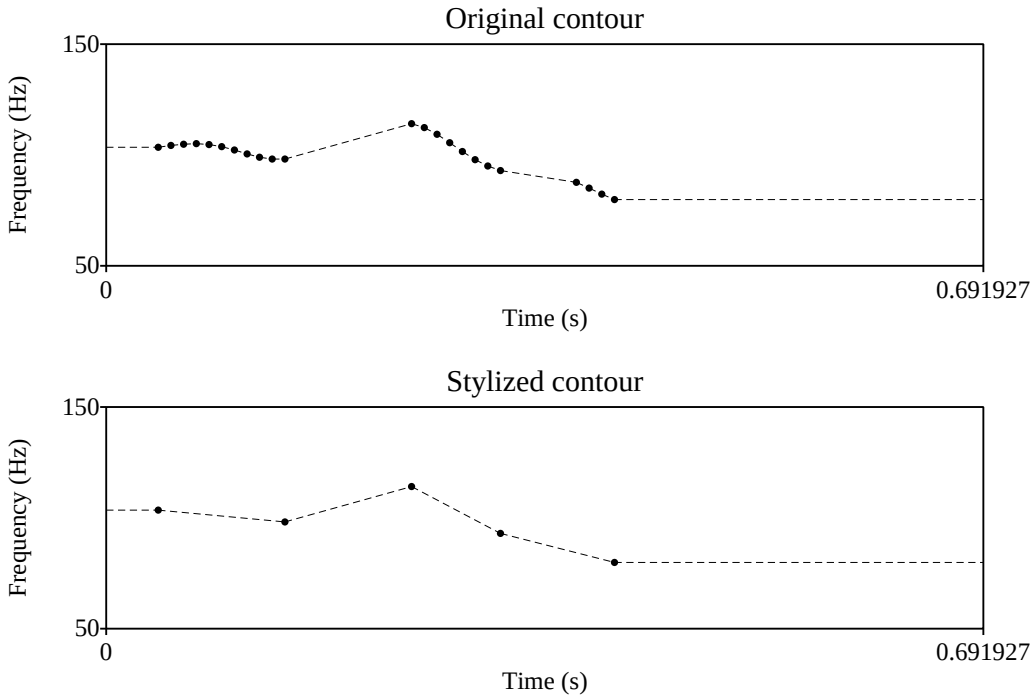

Figure 7.1: Original (upper diagram) and stylized (lower) F0 contours of the word *Helsingfors* 'Helsinki'. The dots show the F0 points, and the dashes how they are connected. In the upper diagram, the dots correspond to the voiced segments of the utterance.

in general tendencies here and we think that the amount of such variation will be roughly the same in each linguistic category.

Another issue is that there is no guarantee that the number of points selected will be the same from word to word. The stylization simply selects the lowest number of points it needs in order to produce a line within the given limit. This means that the number of points used in modelling the original contours potentially may vary with the complexity of each pitch contour. This acoustically grounded variation in the number of features to be predicted is a bit problematic unless the number itself is possible to predict from other linguistic features. Without claiming that we have used the best solution to this problem, we will return to this issue in the model-building described in Section 7.5. The issue was also discussed in Chapter 6.

7.4.2 Pitch extraction

The F0 analysis, the smoothing, and the contour stylization were performed using the functions available in the PRAAT program (Boersma and Weenink, 2003). Some words (about 20) were spoken with a very harsh or whispered voice quality, and they were discarded from the study as it was impossible to calculate F0 in these words. This left 2311 words for the remaining study.

In order to test the ability of the stylization procedure to accurately model F0 contours in real speech, new contours were reconstructed from the stylized data and compared with the originals. The RMSE (*root mean squared error*) between voiced frames of the original and reconstructed contours was 2.68 Hz (0.296 semitones, if calculations are based on logarithmic values) for the 2311 words.

7.4.3 Normalization

Since the corpus contained speech from both male and female speakers, we normalized the actual pitch values by dividing the values of the pitch points by a value calculated by dividing the mean F0 of each word by 100.

The words were also normalized in time by dividing the times of the pitch points by the duration of each word. The values thus obtained correspond to how far into the word a pitch point is, expressed in percentage of the word's total duration.

7.5 Building models

In the previous section we described the extraction of pitch information. Construction of models from this data is not completely straighforward since the pitch feature vectors contain a different number of elements. If one word has been stylized with two points and another in the same category with three points, how can we integrate these parameters in one model? The strategy selected here is: if two different pitch patterns have been stylized with a different number of points in order to keep the stylized contour within the allowed range from the orginal contour, we deem that both pitch patterns are worth using. We thus subcategorize the contours within a word category according to the number of pitch points used in stylizing each contour. Other types of more linguistically motivated partitioning of the data, such as syllable weight, number of syllables in the word or the word's position in the phrase may also prove to be valuable, but this requires a finer analysis than has been performed at the moment and hence has to be saved for a future study.

Following this discussion, all the data was classified according to:

- Order of stylization (the number of pitch points used to stylize the pitch contour)
- Position type of the primary stressed syllable
- Word accent category (Accent 1 / Accent 2)

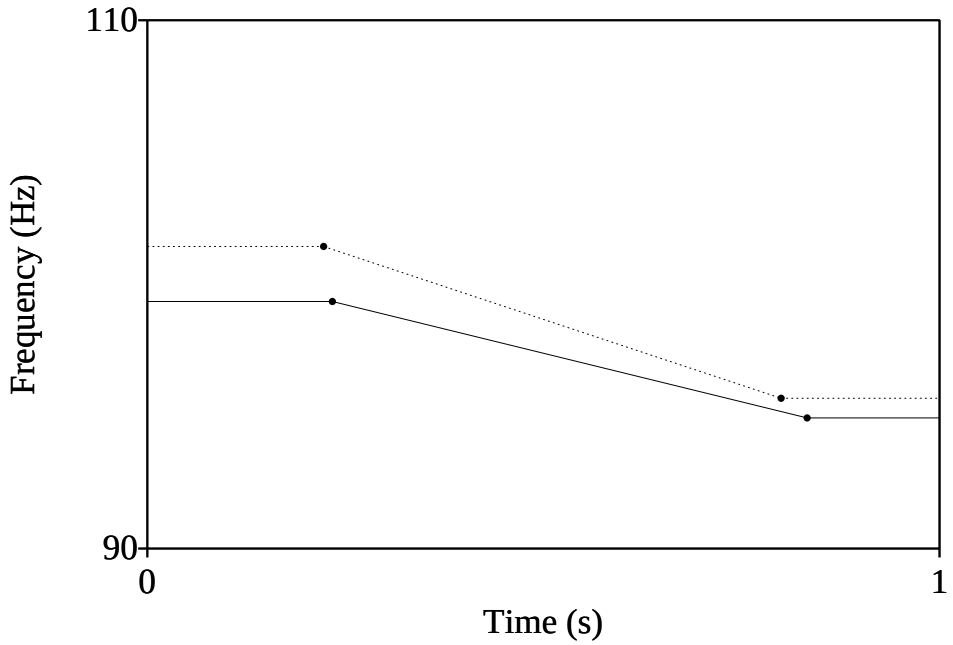

Figure 7.2: Models with stylization order 2, stress type is initial. The full line is the model for Accent 1, the dotted line the model for Accent 2.

All words with the same order of stylization and the same position and accent types were placed in the same group. For each group, the mean time and frequency values for each pitch point were calculated. In order to ensure reliability of the models, only groups with more than 30 occurrences were used. In each group, every fifth word was left out and placed in the test set for that group. In the end, 1974 words were used and the split between training and testing data was: training set = 1591 words, test set 383 words (roughly an 80%–20% split).

Illustrations of the models are shown in Figures 7.2 and 7.3. The figures show the models for words stylized with two and three pitch points and initial stress. They are shown in the normalized format, meaning that the pitch scale is around 100 Hz and the time scale from 0 to 1. The main difference seems to be that the Accent 2 and Compound models generally start from a higher level and fall to a lower level than the Accent 1 model does. This also gives them a steeper shape.

7.6 Evaluation

In order to evaluate the model, intonation contours for the words in the test set were reconstructed from the model. For each word, the model was 'denormalized' using the word's actual length and mean pitch. It would also have been possible to compare the output of the model with the normalized version of the test word, but since we want to use this method for synthesis of real speech we think it is better to

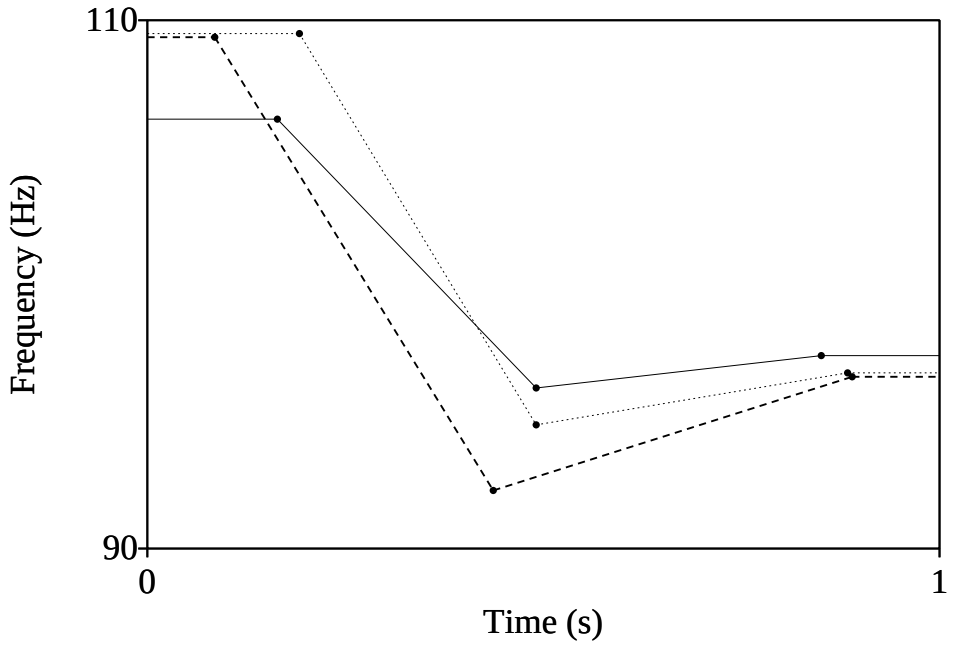

Figure 7.3: Models with stylization order 3, stress type is initial. The full line is the model for Accent 1, the dotted line the model for Accent 2 and the boldface dashed line the model for compounds.

perform the comparison in the 'denormalized' domain. The reconstructed contours were then compared with the originals by calculating the numerical difference between the reconstruction and the original. Note that the numerical measure should only be taken as an indication of how successful a model *might* be. The perceptual impression is always the ultimate test of any model of intonation. No perception test has been carried out so far. Perception tests are quite laborious and not always easily interpretable, and numerical measures can give some indication of whether or not it is worth while to test a model further.

7.7 Results

Table 7.1 shows the mean and median RMSE in all the groups included in the study. The overall measure should be taken with some care, since only the groups with 30 or more words are included, and this measure thus does not account for the other cases.

Table 7.2 shows the results for the different groups. Each error rate is a mean of all the words in that group.

The results show a clear tendency that the lower the order of stylization, the lower the RMSE. Medial and Single stressed syllable types are generally lower than Initial. Only one case of Final is included in the study and this gives the largest

Table 7.1: RMS Errors (in Hz and semitones) between reconstructed and original F0 contours.

	Hz	Semitones
mean	10.6	1.19
median	7.6	0.96

Table 7.2: RMS Errors (in Hz and semitones) between reconstructed and original F0 contours for different groups.

order	Syll. Type	Acc. Type	RMSE (Hz)	RMSE (st)
2	initial	Acc1	5.4	0.65
2	single	Acc1	3.7	0.46
2	initial	Acc2	4.1	0.48
3	initial	Acc1	6.7	0.80
3	mid	Acc1	7.9	0.98
3	single	Acc1	10.0	0.96
3	initial	Acc2	7.8	0.96
3	initial	Comp	6.1	0.79
4	final	Acc1	31.4	3.00
4	initial	Acc1	10.0	1.04
4	mid	Acc1	8.3	0.93
4	single	Acc1	15.0	1.67
4	initial	Acc2	10.4	1.25
4	initial	Comp	9.4	1.00
5	initial	Acc1	12.2	1.33
5	mid	Acc1	13.8	1.46
5	initial	Acc2	18.2	1.82
5	initial	Comp	13.0	1.55
6	mid	Acc1	9.9	1.19
6	initial	Acc2	16.1	2.14
6	initial	Comp	20.4	2.27
7	initial	Comp	14.5	1.65
8	initial	Comp	15.4	1.92
9	initial	Comp	21.3	2.24

error, 31.4 Hz. Acc1 and Acc2 groups generally have lower RMSEs than the Comp groups, but this is probably correlated with the order of stylization. The Comp group is more common in the groups with a higher number of pitch points. For stylization orders 2 and 3, the RMSE is below 10 Hz and 1 st in all the groups.

7.8 Discussion

Many aspects of this study are sub-optimal and work in progress. Focussed words are not distinguished from unfocussed and prosodic context, i.e. position in phrase, pre- or postfocal position, information about the presence of other and adjacent accents, and syllabic information, is not used and the lexical word is not always the most relevant domain for prosody.

Still, we interpret the results as indicative that the stylization method used in the study is able to model intonation patterns accurately. Particularly the words with the lower order of stylization have very low RMSEs.

Comparisons are somewhat hard to perform as most previous studies for other languages than Swedish include everything in a pitch contour. Reported figures on similarities between predicted and modelled pitch contours are often based on the whole contour, even though some parts of the contour are perceptually more important than others. Black and Hunt (1996), for example, using a linear regression method, obtained an RMS error of 34.8 Hz for English and 20.9 Hz for Japanese but this result is for full sentences.

Another problem remaining to be solved, apart from using more linguistic features to categorize the groups, is how the different orders of stylization should be predicted. In a speech synthesis system this has to be guessed from the text processing, and it is not clear at this stage how this should be done. A possible way is to restrict the number of stylization points by selecting them according to some criterion, i.e. only using the first x number of points after some specific event like a CV boundary. Such a method is used in Chapter 8. Another possibility is to encode the number of stylization points for each word in a dictionary.

In a way, the patterns resulting from calculating the means are quite similar to vector quantization (VQ) in that they represent a number of distinct pitch patterns. Using VQ in intonation modelling is a promising approach, as shown in Möhler and Conkie (1998) and Syrdal et al. (1998).

7.9 Conclusions

This study has shown that a model that uses stylization of pitch contours for accented content words in Swedish and the linguistic features word accent type and stress position type is able to predict pitch contours that numerically are quite similar to natural ones. The method works particularly well for less complex intonation patterns.

Chapter 8

Automatic classification of word accent, focus and dialect type

8.1 Introduction

This chapter deals with the automatic classification of dialects, word accents and focal accent in Swedish. We used compounds and disyllabic words from the speech material of the SWEDIA 2000 dialect project in order to build statistical models for the prediction of these intonational phenomena. The model is intonation-based and uses different parameters extracted from F0 contours of the words.

It has long been known that intonation is a major source of variation in the dialects of Swedish. An important factor causing this variation is found in the intonation data collected by Meyer (1937, 1954) where it is clear that the temporal alignment of turning points in the F0 contour differs between dialects. Further evidence was brought forward in different experiments by Bruce and Gårding (1978) and Bruce (1983), which showed that the perceived dialect type of an utterance could be varied by means of resynthesis with different F0 contours. Bruce (p.c.) has also demonstrated that dialect type may be recognized from the laryngograph (Fourcin and Abberton, 1971) signal of an utterance only. In such a signal, all vowel and consonant identity is neutralized and this indicates that listeners are able to use prosody, in particular F0, as a cue to dialect. House (1990) and House and Bruce (1990) have dealt with automatic prosody recognition. Based on the performance of a human expert's analysis of F0 contours, they develop a set of rules for the classification of unknown F0 contours.

In another study, Malmberg (1955) demonstrated that F0 was more important for distinction of word accent than intensity and duration. The importance of different F0 patterns has later also been shown by, for instance Bruce (1977).

Focus detection has been studied by Sautermeister and Lyberg (1996). They extract F0 from the speech signal of sentences and perform a normalization based on declination. A later study by Heldner et al. (1999) examines focus detection in controlled material. They use measurements of intensity and spectral tilt to detect focus in three-word sentences. Even though they do not use F0 patterns as such, they use a measurement of F0 in the spectal tilt measurements. Their best method is based on intensity and finds the correct focussed word in 72% of cases. Eklund and Lyberg (1995) sketch a model for the extraction of accentual and dialect information in Swedish. Speech recognition systems may recognize accents and dialects by using a string of recognized segments and performing comparisons between this string and F0 patterns.

The experiments presented here are based on the same basic idea: intonation, as manifested in the F0 contour, plays a part in the prediction of certain features of dialectal and accentual variation and it is possible to automatize the recognition of some aspects of this relationship.

A possible application for dialect recognition is in voice response systems, which need to be able to cope with dialectal variation in order to maximize the number of potential users. Identification of accent type may restrict lexical possibilities and therefore facilitate lexical search and access in automatic speech recognition systems. In fact, for pure tone languages, prosodic recognition may be crucial. Another use is in speech research: an automatic accent detector would faciliate the annotation and development of databases for use in research within speech technology and phonetics.

8.1.1 Outline of the chapter

We start out with a small experiment (Experiment I) based on differences in the realisation of compound accent in Southern Swedish, where we investigate the distribution of word accent in this region. We use the same material to develop a method based on F0 parameterization with the aim of distinguishing the two word accents acoustically (Experiment II). We then present two larger studies on the classification of accents and dialect types with material covering, in Experiment III, a larger area in the south of Sweden (the region called Götaland), and in Experiment IV, the whole Swedish-speaking language area. Experiment III and IV are rather similar. The major difference is that Experiment IV uses material from the whole Swedish-speaking language area. Experiment III, apart from the results, may be seen as a test of the methodology with a smaller amount of material than the full study. The reason that we performed the study with an smaller and geographically incomplete data set was mainly that this material was available to us prior to the full Swedish material, but also because the dialects in this region cover four of the five

major groups described in the dialect typology in Gårding (1977)[1], and hence still is quite representative of the dialectal differences in Swedish. Finally, in Experiment V, we perform a study where we derive the most useful intonational features (of the intonation models under consideration) in each dialect for distinguishing the word accent types and the nonfocal/focal words.

8.2 Experiment I: Compound accent patterns in some dialects of Southern Swedish

This is a study about the accent pattern of compound words in the Southern Swedish variety. This pattern differs from Standard Swedish in that the first element of the compound may retain Accent 1, whereas the standard dialect invariably gets Accent 2. Bruce (1974) showed that this resistance to shifting the word accent varied according to area and type of compound. The present study deals with five dialects of Southern Swedish using the recently collected material from the SWEDIA 2000 project (Bruce et al., 1998).

In Swedish, two degrees of intonational prominence are generally recognized: accent and focus (see Table 5.1). As has been described in earlier chapters, Swedish has two word accents, **Accent 1** (or acute) and **Accent 2** (or grave), that are realized phonetically by different F0 movements in relation to the stressed syllable. The word accents are primarily lexical, and they both co-occur with semantic or phrasal focus, resulting in a focal accent that is realized by a slightly different pattern of F0 movements. The word and focal accent patterns are also one of the major characteristics of the differences between Swedish dialects. A case in point is compound words, where the accent pattern is different in Southern and Standard Swedish. Compound formation is highly productive in Swedish and occurs frequently in everyday speech as well as in more formalized situations. For a more detailed account of non-accentual properties of Swedish compounds, see e.g., Liljestrand (1993).

8.2.1 Compound words and compound stress

By compound words we mean words that have the compound stress pattern, i.e., one primary stressed syllable and one secondary stressed syllable. The primary stress usually falls on the leftmost element of the compound, and the secondary stress on the rightmost element. These words include: (1) real compounds (combination of two or more simple words), (2) derivatives (formed by adding stressable affixes) and (3) formal compounds (words that have the compound stress pattern but cannot be analysed into independent morphemes)

[1] This is true if we put the Gotland dialects in the 1B (Dalarna) group, which is prosodically viable.

(1) *järnväg*, *stoppskylt*, *trafikljus*
'railroad, stop sign, traffic light'

(2) *mångkulturell*, *trippelseger*, *läraktig*
'multi-cultural, treble victory, ready (or willing, apt) to learn'

(3) *abborre*, *äventyr*, *paradis*, *ingefära*, *rädisa*
'perch, adventure, paradise, ginger, radish'

Morphological compounds

Some words may consist of two or more stressable morphemes and are therefore compounds or derivatives morphologically, but they do not have the compound stress pattern. Instead they receive lexical phrase stress, which is right dominant (stressed on the last element). This occurs when the suffix attracts the primary stress (4). Furthermore, some geographical suffixes (5) are able to repel secondary stress in some compound formations, as well as the suffixes in (6).

(4) *-al*, *-inna*: central, lärarinna
'centre (or central), teacher (fem.)'

(5) *-land*, *-stad*, *-bo*: Finland, Ystad, Åbo
'Finland, Ystad, Åbo'

(6) *-dag*, *-bär*, *-gård*: måndag, vinbär, skärgård
'monday, currant, archipelago'

The morphological compounds in (5) or (6) get one stress only, either on the first or last element. Note, however, that there is a regional variation regarding these suffixes. In some varieties of Swedish they may get the compound stress pattern.

8.2.2 The accent pattern of compound words

In Standard Swedish, compound words receive Accent 2 on the element starting with the primarily stressed syllable. This means that words having Accent 1 (or lacking Accent 2) in isolation are assigned Accent 2 in compounds. This rule applies almost indifferently; only in cases like (5) and (6) does the first element retain its word accent. In Southern Swedish, the pattern is different. Accent 1 words acting as first elements in compounds may resist the accent shift, with the result that the compound word has Accent 1. Bruce (1973) states the rules governing this resistance as in (7)-(8)[2]. Compounds with Accent 2 are illustrated in (9).

[2]Word accents are indicated using the traditional symbols: ´ for Accent 1 and ` for Accent 2. The symbols appear above the vowel in the stressed syllable except when this vowel already contains a diacritic (ring or diaeresis). In this case the accent symbol precedes the vowel letter.

Exceptions to (9b) do exist, e.g. colour adjectives and numerals: *svártvitt* 'black and white', *trérummare* 'three-room [apartment]'.

(7) Words with a monosyllabic first element receive Accent 1 when:

(a) directly followed by an unstressed syllable in the second element:

gáspedal, *stórpublik*, *cúpfinal*
'accelerator [pedal], large audience, cup final'

(b) directly followed by infix s:

tv'ångsmatning, *bórdsplacering*
'force-feeding, placing at the table'

(c) the second element is structurally and semantically more related to the following elements than the first:

bárndaghem, *bússhållplats*
'daycare centre, bus stop'

(8) Words with a polysyllabic first element receive Accent 1 when:

(a) finally stressed:

totállösning, *finálbetoning*, *adrésslapp*
'full solution, final stress, address label'

(b) with penultimate (or earlier) stress and having Accent 1:

bándyboll, *narkótikabrott*
'bandy ball, drug crime'

(9) Words receive Accent 2 when the first element of the compound:

(a) already has Accent 2:

gàmmalmodig, *j` ättefin*
'old-fashioned, terrific'

(b) is monosyllabic and is directly followed by a stressed syllable:

màtsal, *stòrstad*, *flèrstämmig*
'dining-room, big city, polyphonic'

In a follow-up study, Bruce (1974) checks the validity of these rules using informants from the dialect areas of Malmö, Kristianstad and Halmstad (all in Southern Sweden). Most of them were younger speakers, university students at the time. The cases in (9) behaved as expected with an Accent 2 pattern. The

results for the other types are summarized in Table 8.1. Based on these results, an implicational relationship ('ranking hypothesis') is suggested where the types are ordered on the basis of their ability to resist Accent 2 in compounds. The ranking is (8b) > (7b) > (8a) > (7a) > (7c).

Table 8.1: Occurrence of accent shifts in different dialects in Bruce (1974)

Type	**Example**	**Malmö**	**Kristianstad**	**Halmstad**
(7a)	gáspedal	no	yes	yes
(7b)	tv´ångsmatning	no	no	yes-no
(7c)	bússhållplats	no	yes	yes
(8a)	finálbetoning	no	yes(no)	yes
(8b)	bándyboll	no	no	no

8.2.3 Aim of present study, research questions

Since a rather considerable amount of time has passed since this study, and one common phenomenon of dialects is adaptation to the standard language, it would be interesting to see what the situation is today. The recent research project SWEDIA 2000 aims at collecting speech data of many different dialects of Swedish. This elicited material contains compound words, and thus allows us to investigate the matter. Unfortunately, the material in Swedia has limitations, so all the compound types can not be tested. It is, however, possible to examine the accent pattern of type (8b), which has the highest ranking in Bruce's hierarchy. This pattern is therefore most likely to still be able to resist Accent 2 and is perhaps the most interesting to examine.

The first question we shall try to answer in this study is: How does the realization of compound accentuation vary depending on age, gender and dialect? The more general second question is how the realization of compound accentuation today relates to the findings of Bruce (1974). The third question is related to the acoustic aspects of the accent distinction and automatic recognition of word accent. Is it possible to formalize the F0 patterns of the different word accents so that it is possible to write rules for word accent classification?

8.2.4 Data and analysis

In this study, we used material collected within the SWEDIA 2000 dialect project (Bruce et al., 1998). In the Swedia material, there are recordings from more than 100 villages and towns, mainly located in Sweden but also a few in Finland, where Swedish is spoken in some areas. Twelve subjects are recorded in each village, distributed across age and gender. Three subjects in each of the following groups were recorded.

(a) younger (20-30 years of age) men (YM)

(b) younger women (YW)

(c) older (50-75 years) men (OM)

(d) older women (OW)

For this study, we selected five[3] villages from the Swedia material from the southernmost part of Sweden. They were selected because they are situated close to the dialect areas used in Bruce (1974). For maps, see Appendix A.

(1) Bara, in southwestern Skåne (close to Malmö)

(2) Löderup, in southeastern Skåne (southeast of Malmö)

(3) Norra Rörum, in central Skåne (approximately halfway between Malmö and Kristianstad)

(4) Broby, in northeastern Skåne (close to Kristianstad)

(5) Våxtorp, in southeastern Halland (close to Halmstad)

For every group of speakers, several compound utterances were included. However, only the following utterances were potential Accent 2 resisters:

(I) *endollarsedel* (one_dollar_bill), a compound word consisting of the words *én* (one), *dóllar* and *sédel* (bill). This is a case of an exception to (9b), having a monosyllabic numeral as first element. This compound word has primary stress on the syllable 'en' and secondary stress on the syllable 'se'.

(II) *femtiolapp* ("fiftier", a bill of fifty crowns), a compound word consisting of the words *fémtio* (fifty) and *lápp* (slang meaning note). This is a type (8b) compound and has primary stress on the syllable 'fem' and secondary stress on the syllable 'lapp'.

In the recording sessions, each phrase is uttered a number of times, giving several repetitions of each word for each informant. The phrases were not read from any written versions, but elicited by the interviewer by showing the informant notes with symbols for each word. The recordings were made in the informants' homes using a portable DAT-recorder and care was taken to avoid background noise. In almost all cases the recordings are of very high quality.

[3]A sixth place in the southern area, Össjö, was recorded later that this analysis

Some utterances had to be discarded either because the intensity was too low, or because the wrong target word was said, for instance, some speakers mistook the numeral for an indefinite article and said *en dollarsedel* 'one dollar-bill' (two-word phrase). In all there were 116 utterances.

The utterances were classified as having Accent 1 or Accent 2 by the author and a highly trained phonetician. Both come from the Southern Swedish language region and were familiar with the distinction in the realization of the different word accents in this dialect area. Focus intonation and phrase accents were abstracted away from, as this does not influence the type of word accent. The utterances were first analysed by ear and in some unclear cases through analysis-by-synthesis after F0 manipulation. This allows the experimenter to change the F0 contour and do a resynthesis with the new contour. In this way the word accent can be changed from one to the other and this is helpful in determining which accent a word has. However, in > 90% of the cases auditory analysis was sufficient. This analysis was only performed by the first listener. All in all, 232 judgements were made.

The different accent patterns are illustrated in Figure 8.1. The figure shows focused versions of the utterance *endollarsedel* for one speaker in Bara and one speaker in Broby. The Bara speaker (an Old Man) has Accent 1 and the Broby speaker (also an Old Man) has Accent 2. The Bara speaker also has a 'continuation rise' at the end of the utterance, but this is unrelated to the accent pattern. Note that the alignment of the pitch accent fall with the onset of the stressed vowel in *en* is much later in the Accent 2 realization.

8.2.5 Results

The results of the accent classification are presented in Table 8.2. The two listeners made almost identical judgements and therefore the results are grouped. For all villages except Våxtorp, there are very small differences depending on the word, of on the speakers' gender and age. The only major variation is geographical. From Table 8.2, it is clear that all of the speakers in Löderup and Norra Rörum use Accent 1 in both words. In Bara, almost all speakers use Accent 1, except for one of the Old Men. In Broby, the Accent 2 pattern dominates (also with the exception of one of the Old Men). Finally, in Våxtorp, the picture is more variable. Here, all speaker groups exhibit some variation: speakers seem to use either the Accent 1 pattern or the Accent 2 pattern. An interesting difference is between Young Men and Young Women, where the former tend to use Accent 1 more and the latter Accent 2 more.

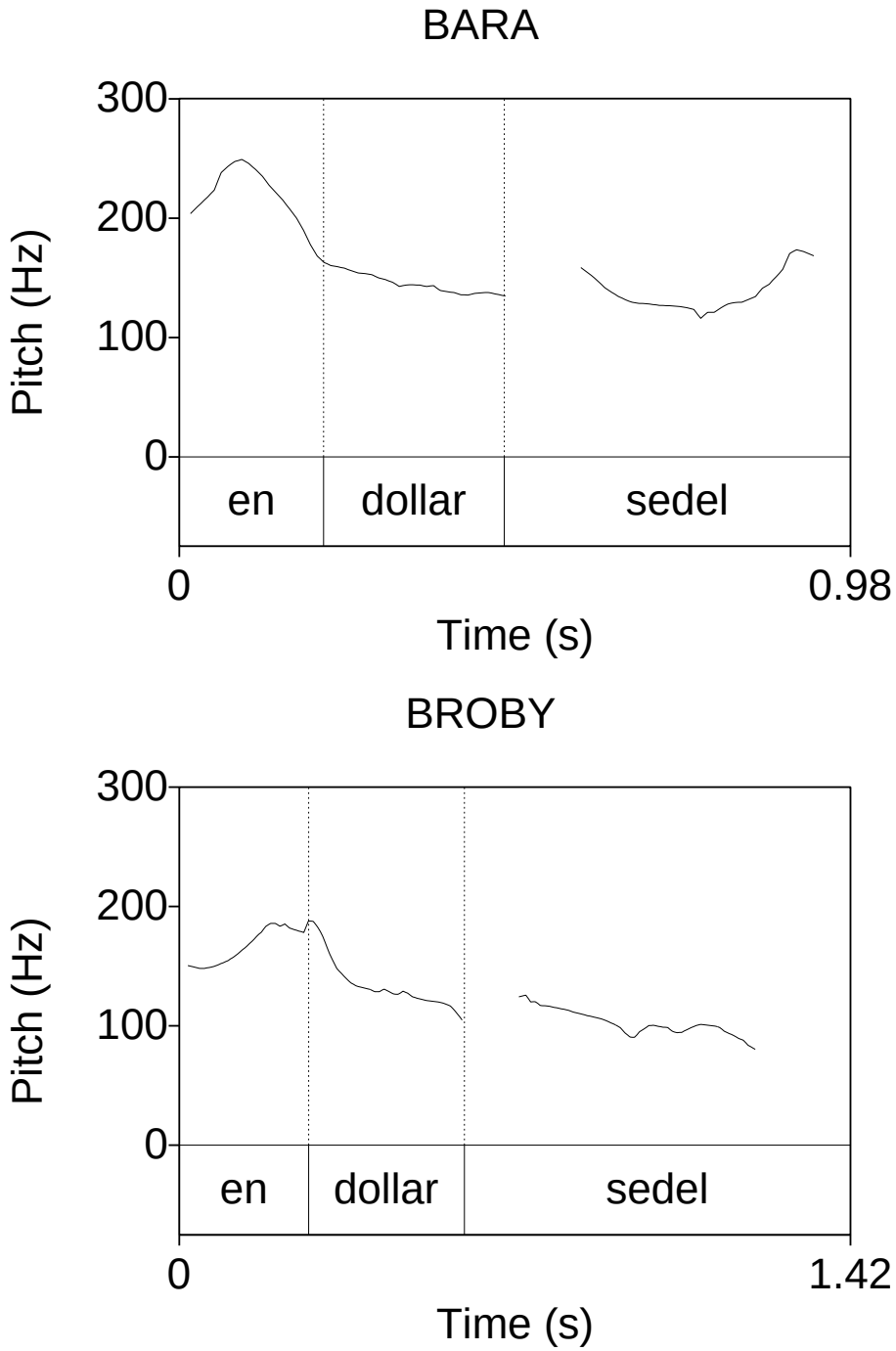

Figure 8.1: F0 tracks of the utterance *endollarsedel* in Bara and in Broby. The utterances are segmented into the different morphological parts of the words.

8.2.6 Discussion

This study indicates (by way of extrapolation) that the accent shift resistance is still prevalent in Bara, Löderup and Norra Rörum, while it is virtually nonexistent in Broby. If we follow the ranking in Bruce (1974), we see that even the highest

Table 8.2: Listening results for two listeners. Number of occurences of Accent 1 for each village, word (I or II), gender and age and the number of analysed words in each group.

		Word (I)	**Word (II)**
Bara	OM	6/8	5/8
	OW	4/4	6/6
	YM	4/4	4/4
	YW	6/6	6/6
	Tot.	**20/22**	**21/24**
Broby	OM	1/6	1/6
	OW	0/6	1/6
	YM	0/6	0/6
	YW	0/6	0/6
	Tot.	**1/24**	**2/24**
Löderup	OM	4/4	6/6
	OW	6/6	6/6
	YM	4/4	4/4
	YW	6/6	6/6
	Tot.	**20/20**	**22/22**
Norra Rörum	OM	6/6	8/8
	OW	6/6	6/6
	YM	4/4	6/6
	YW	6/6	6/6
	Tot.	**22/22**	**26/26**
Våxtorp	OM	4/6	5/6
	OW	4/6	4/6
	YM	5/6	5/6
	YW	1/6	1/6
	Tot.	**14/24**	**15/24**

ranked compound type (8b) gets Accent 2 in Broby. A prediction from this would be that the resistance to shifting word accent is lost in all types. Some support comes from the other word *endollarsedel* that follows (9b) and fails to follow the exception pattern it has in the other dialects. In Våxtorp the resistance is unstable. Both words are uttered with both accent patterns.

The fact that Broby does not show resistance to word accent shifting raises the question of where the geographical border of the accent shift is. Broby is not very far to the north of Kristianstad, and south of Halmstad, both of which retained Accent 1 in the top-ranked (8b) pattern in Bruce's data. The border of the accent shift blocking may thus have moved southwards as a result of adaptation to the

standard dialect of Swedish, but the lack of differences between generations belie this. Växtorp, on the other hand, seems to be locked right on the accent shift border.

Except for in Växtorp, age and gender seem to play minor roles. There are a few exceptions in the Older Men class, but this may also be idiolectal. There was a small variation between the two listeners. However, we still think that the regional differences are adequately reflected in the results.

8.3 Experiment II: F0 based word accent classification

There are many studies on the topic of automatic prosodic feature extraction of accentuation that concentrate on the shape of the F0 contour. The general idea behind these studies is to use the most prominent rises and falls in the F0 contour in order to define features such as pitch level and the timing of turning points (often related to vowel onsets) and then develop a rule system for classification. The rules may be developed intuitively (House and Bruce, 1990) or be based on some sort of machine learning technique (Dusterhoff et al., 1999; Streefkerk, 2002; Sakurai et al., 2002).

8.3.1 Data and analysis

Following this methodology, we used the same material as described in the previous section (Section 8.2) and attempted to formalize the difference in the acoustic realization of the two word accents. The goal of this experiment was to develop a system that would be able to predict the type of word accent based on the F0 contours. In order to parameterizise the F0 patterns we used the stylization procedure described in Section 6.7.2. The threshold for allowed deviation from the original contour was set to 2 semitones. In Figure 8.2 we show an example of the stylization procedure for the two utterances in Figure 8.1.

It was rather suitable to use the utterances from Section 8.2, as the word *endollarsedel* begins with a stressed vowel and the word *femtiolapp* has a voiceless fricative before the stressed vowel. The onset of the stressed vowel was therefore trivial to locate as it almost always was the first registered F0 point in the whole utterance and very little or no segmentation was necessary. The stylization points were always counted starting from the vowel onset.

We measured the F0 level and the timing of the three first resulting stylization points, and then we used the differences between the timing and the level values, respectively. As the first measurement point was always at the vowel onset this means that our temporal data are relative to the vowel onset. In order to be able

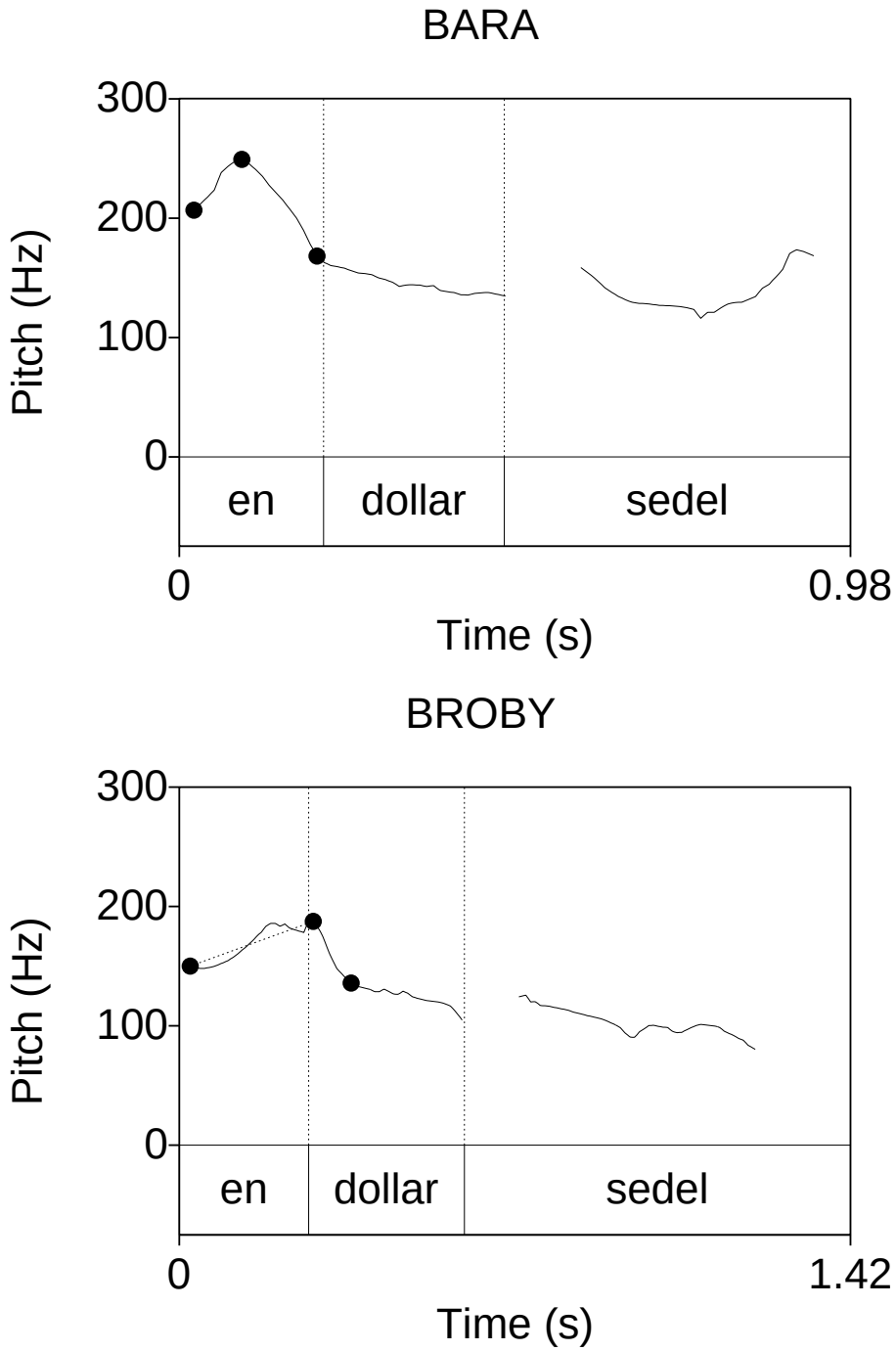

Figure 8.2: Stylizations of the F0 tracks in Figure 8.1. The black circles are the stylization points. The dotted line between the two first points in the BROBY graph does not deviate from the original by more than 2 st.

to compare men and women the F0 level values were measured in semitones. To summarize, the features we used were:

- **t2-t1** Timing difference between the two first points.

- **v2-v1** Frequency difference between the two first points.
- **t3-t1** Timing difference between points 3 and 1.
- **v3-v1** Frequency difference between points 3 and 1.
- **t3-t2** Timing difference between points 3 and 2.
- **v3-v2** Frequency difference between points 3 and 2.

These measurements were performed for all 116 utterances. Part of the data is presented in the scatterplot in 8.3. There we see the positions of the second measurement points related to the first measurement points for every word in the material. The position of a given point is thus determined by its distance in time (s) and frequency (st) from the vowel onset. A frequency difference of 0 thus means that the F0 contour is flat until the first stylization point after the vowel onset, whereas positive differences indicate a rising contour and negative differences are the result of a falling contour through the vowel.

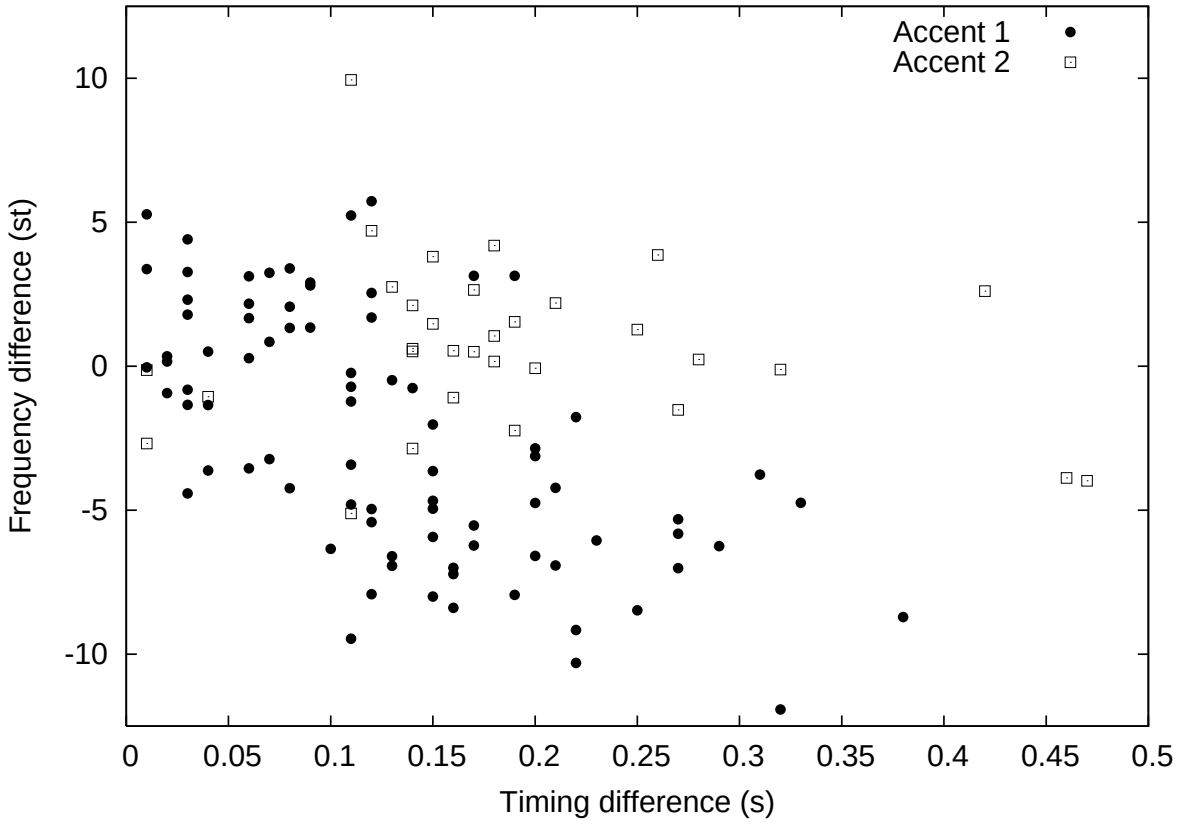

Figure 8.3: Distribution of the data.

We can see the distribution of the different word accents in the plot. Even though there is some overlapping between the regions (as well as a few outliers) it

seems to be possible to formulate some rules that will distinguish between most cases. Using the CART technique (described in Chapter 4), the optimal rules and their ordering can easily be found for this dataset[4]. It appears that we can apply a very simple set of rules and achieve a very good discrimination. The rules are:

1. If the timing difference between point 2 and point 1 is larger than 130 ms, then classify the word as Accent 2.

2. If the F0 level difference between point 2 and point 1 is less than -2.552 st, then classify the word as Accent 1.

8.3.2 Results

The rules are visualized in Figure 8.4, which basically is the same plot as Figure 8.3, but with the dashed lines indicating where the rules apply. Words whose first stylization point after vowel onset is in the area to the left of the vertical dashed line (at 130 ms) and below the horizontal dashed line (at -2.552 st) are classified as Accent 1, the others as Accent 2. These rules classify 100 out of 116 words correctly (86.2%). This is above the performance of a baseline classifier. The baseline classifier would classify all cases as the most frequent type, in this case Accent 1. There are 83 cases of Accent 1 out of 116 words. This would imply a correct result in 71.5% of cases.

It is possible to improve the classification by adding more rules that use the other features as well. As this results in a more complex rule structure, we cannot visualize it directly on the data. However, it is possible to produce a tree depicting the rules and their ordering. It is shown in Figure 8.5. This tree classifies 107 cases correctly (92.2%). The questions at the nodes should be interpreted as yes-no questions, where the left leaf is the branch taken by a 'yes' answer to the question and the right branch by a 'no' answer. The tree is walked from top to bottom until a leafless (terminal) node is reached and then the content of that node is returned as the output of the tree. Thus, if the timing difference between the two first points (t2-t1) is less than 0.130 s, then the output of the tree is '1', meaning Accent 1. Note that our simple rules are at the top of the tree. While these features still are the most important ones, the other features help to improve the performance of the rule system.

8.3.3 Results for recognition of unseen data

So far, we have been discussing possible ways to characterize the data, but we have not yet tested the rules on unseen data, so we can not yet claim that the rules may be

[4] With such a small dataset, it would not have been very difficult to extract these rules by hand.

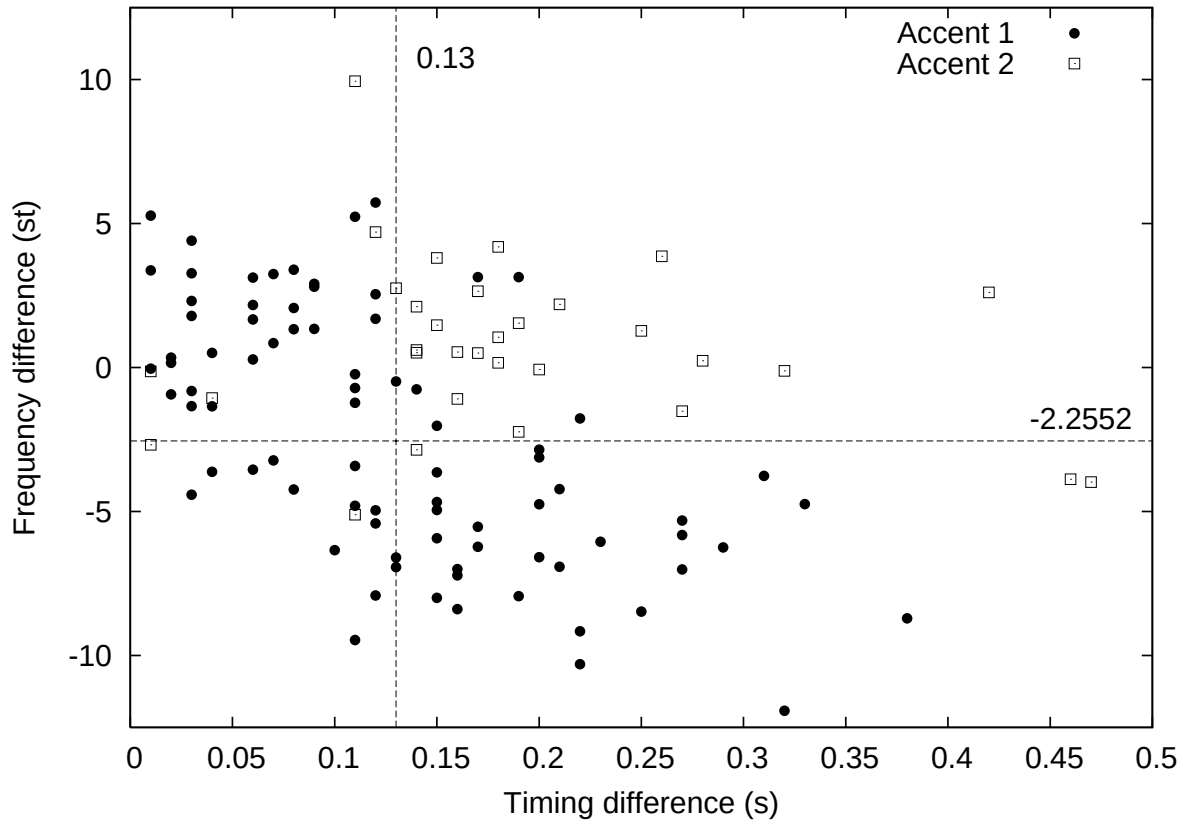

Figure 8.4: Visualization of a simple set of rules for accent distinction.

suitable for accent recognition. In order to perform such a test, we made an 80-20 split of the data, putting 93 cases in the training set and 23 in the test set. We then built a CART tree using only the training set and used it to classify the data in the test set. This resulted in a really good performance: 21 out of 23 (91.3%) cases were correctly classified.

8.3.4 Discussion

It may seem surprising that the classification performance for unseen data equals that achieved when using all the data. It could be that the problematic cases (those on the 'wrong side of the line' if we look at Figure 8.4) end up in the training set, and therefore the errors do not show up in the test set. However, if we switch the training and testing sets (thus using a very small training set and possibly allowing for many potentially erroneous cases in the test set) it is still possible to achieve an impressive 83.9% correct. We take this as an indication that there is a genuine difference in the acoustic realisation of the two word accents and that it is possible to capture the difference using the method described in this section.

As this is a rather limited material coming from a rather homogeneous dialect area we did not perform any further studies on this material. Instead, we ex-

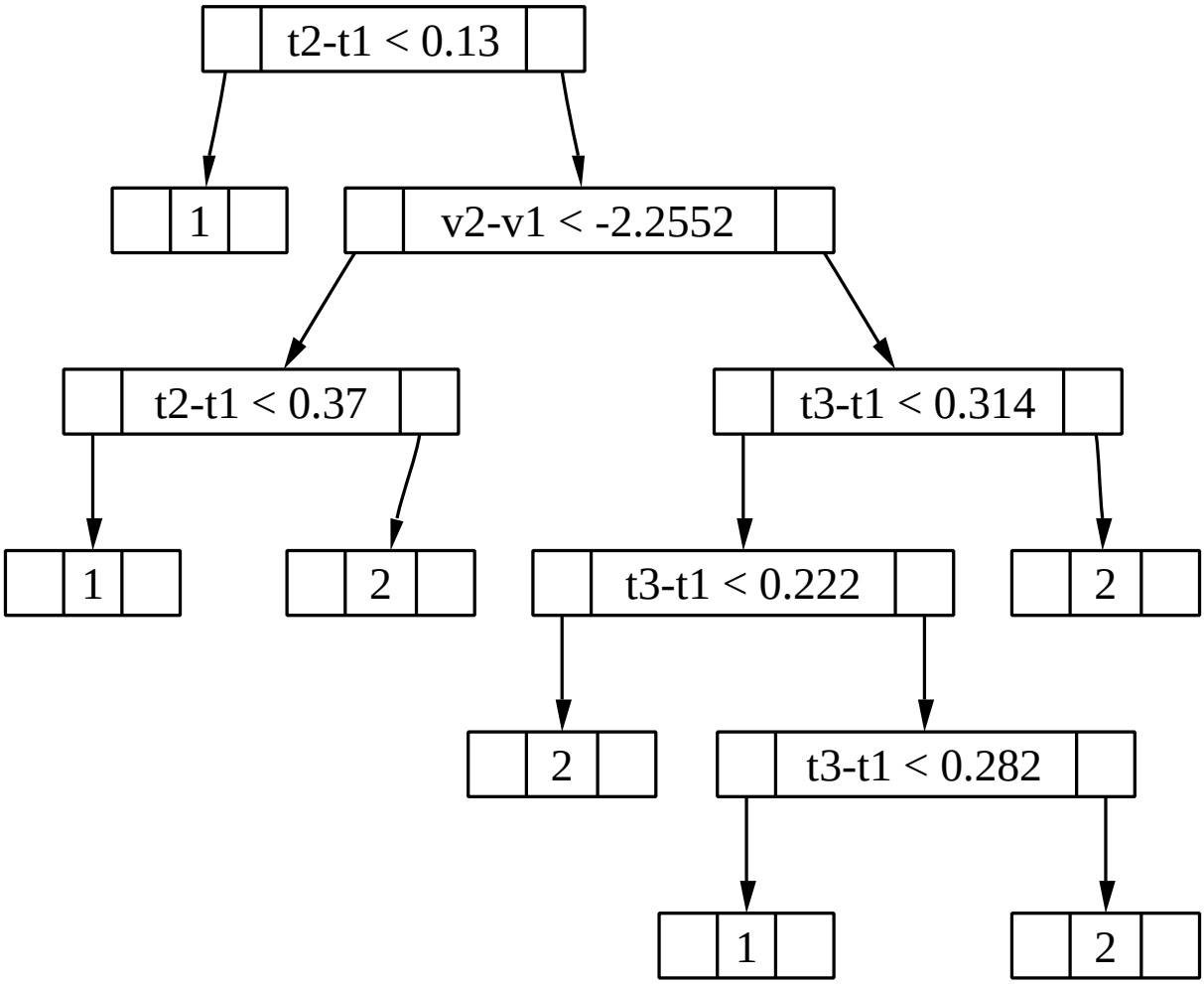

Figure 8.5: A tree with rules for accent prediction.

tended the study to a larger material from the SWEDIA 2000 database and tested other analysis methods as well, in order to predict not only word accent, but also prominence level and dialect type. This will be the topic of the next sections.

8.4 Experiment III: Classification of word accent, focus and dialect type, material from Götaland

In the previous sections we have seen that it is possible to distinguish between the two word accents in material from one part of Sweden on the basis of intonation parameters. However, the realization of word accent type in Swedish varies with geographical location. A full-fledged accent classifier should be able to deal with word accents in all dialects. We therefore now extend the previous experiments with a study performed with more material from a larger geographical area. Since

we add more dialects, this also adds the possibility of testing the predictability of dialect type. Furthermore, the realisation of focussed and nonfocussed words is different from dialect to dialect. The question we would like to answer is: to what extent can acoustic properties, like the timing and level of turning points in the F0 contour, be used to recognize dialect, type of accent and prominence level correctly?

8.4.1 Material

The material was taken from another part of the Swedia database (see Section 8.2). The material we will use here consists of the words *dollar* (Accent 1) and *kronor* (Accent 2) spoken in phrases like 'tio dollar' or 'tjugo kronor', where either the numeral or the currency word is given phrase focus. Thereby we get both focal and non-focal versions of both A1 and A2 words. This material contains recordings from all provinces between Skåne in the south up to Dalsland, Östergötland and Gotland in the north. All in all, we have more than 100 speakers from the ten southernmost provinces of Sweden (see A for a map).

Labelling

The material was first labelled on the word level, locating the start and end position of each word in the whole recording session of an informant. In this process, word accents and phrase foci were also indicated. The material was then phonetically labelled using a semi-automatic method. First, a phonetic transcription was derived from the orthographic transliteration through automatic dictionary look-up. This transcription was then time-aligned with the speech signal using a DTW technique where the natural speech signal is compared with a speech signal synthesized from the phonetic transcription. The technique is described in Malfrère and Dutoit (1997) and has also been used by Black and Lenzo (2003) for segmentation of diphone databases. An evaluation of the technique showed that approximately 25% of the segments are placed less than 5 ms from the position where a human transcriber would place them and another 25% is within 10-15 ms. Since the technique is still not perfect, manual post-processing, where the segment boundaries are adjusted, was performed. In this way, the temporal location of segment boundaries, most notably the important information about vowel onset in the stressed syllable, is obtained for each word. From this material we used the group 'Older Men'. Due to reasons of priority within the project, the labelling of these was the first to be completed and ready for use.

8.4.2 Acoustic analysis and parameterization

Three different sets of parameters were extracted from the words in the material:

1. **FALL** method. Time and F0 values of the first fall where the beginning of the fall comes before the end of the vowel in the stressed syllable. The temporal locations of the turning points were determined by a stylization method (see below). Following Bruce (1977) and Bruce and Gårding (1978), the method assumes that the perceptually relevant cue for accent is a fall somewhere near the vowel in the stressed syllable. Currently, the method is 'greedy'; i.e. it tries to identify the longest possible fall in the whole contour. This means that sometimes the end points may not be the ones a human analyser would choose, since the F0 contour may continue to fall after the most relevant part of the fall. There is thus room for future optimization of this method. The temporal measurement points were related to the vowel onset and to the vowel length in the word. We will use the following notations for these features:

Feature	Description
`ft1_vs`	time at start of fall - vowel onset
`ft1_vs_vdur`	time at start of fall - vowel onset, related to vowel length
`ft2_vs`	time at end of fall - vowel onset
`ft2_vs_vdur`	time at end of fall - vowel onset, related to vowel length
`fall_dur`	duration of fall
`fall_height`	height of fall
`fall_start_st`	frequency at start of fall
`fall_end_st`	frequency at end of fall
`fdur_vdur`	duration of fall, related to vowel length
`fhei_fdur_vdur`	`fall_height`, related to `fdur_vdur`

2. **POINT** method. F0 level at the onset of the vowel in the stressed syllable. Time and F0 level of the two first stylization points (see below) after the vowel onset. This method makes no explicit assumption about the direction of F0 after the vowel onset, but only tries to capture the acoustic features of the turning points directly following it. This method was used with some success in Section 8.2 above in distinguishing between Accent 1 and 2 in material from Southern Sweden. The temporal measurement points were related to the vowel onset and to the vowel length in the word. We will use the following notations for these features:

Feature	Description
t1_vs	time at first point - vowel onset
t1_vs_vdur	time at first point - vowel onset, related to vowel length
t2_vs	time at second point - vowel onset
t2_vs_vdur	time at second point - vowel onset, related to vowel length
point0_st	frequency at vowel start
point1_st	frequency at first point
point2_st	frequency at second point
point_dur	temporal difference between point 1 and point 2
point_height	frequency difference between point 1 and point 2
stdur_vdur	point_dur, related to vowel length
sthei_stdur_vdur	point_height, related to stdur_vdur

3. **TILT** method. These values are based on the Tilt model by Taylor (2000). In this model, each intonation event is characterized by continuous parameters representing amplitude, duration and tilt (the shape of the event). The parameters are extracted using the *tilt_analysis* program distributed with the Edinburgh Speech Tools (see also Chapter 6). The temporal measurement points were related to the vowel onset and to the vowel length in the word. Note also that the search space was limited to the vowel boundaries. We will use the following notations for these features:

Feature	Description
tilt_f0	f0 level of tilt event
tilt_amp	amplitude of tilt event
tilt_dur	duration of tilt event
tilt_tilt	tilt of tilt event
ttime_vs	time at tilt event - vowel onset
ttime_vs_vdur	time at tilt event - vowel onset, related to vowel length
tdur_vdur	duration of tilt event, related to vowel length

Note that we do not use the absolute temporal measurement points when predicting word accents at all. This is because these would show a large variation between the A1 words ('dollar') and A2 words ('kronor'), simply due to the fact that in the A1 words the stressed vowel is preceded by a 'd', whereas in the A2 words the stressed vowel is preceded by 'kr'. The difference between the means of the temporal locations of the vowel onsets in the two words is statistically significant (z=53.78, p<.0002). It is thus not possible to ascribe any accentual differences to the absolute temporal measurement points.

For Focus and Gårding type (see Section 8.4.5) our tests included the absolute points as well. For Focus prediction, the results were not improved, and the figures

presented below are based on relative measurement points. For Gårding type we only tested the case with absolute points, so all results for Gårding type are obtained when including these.

Pitch analysis

The pitch contour used for methods 1 and 2 was obtained by the pitch analysis algorithm by Boersma (1993), that is implemented in Boersma's PRAAT (Boersma and Weenink, 2003) program. This is integrated with the stylization method used in these methods. For method 3 the method by Bagshaw et al. (1993), distributed with the Edinburgh Speech Tools, was used, as this implementation is integrated with the extraction of the parameter set in question (the Tilt parameters). In order to avoid octave jumps etc, the raw pitch data underwent inspection and, where necessary, a reanalysis with adjusted low-pass and high-pass filter settings was performed. This reanalysis was performed before the parameterizations as there are limits to how much and what kinds of post-processing of the F0 contours that may be performed in an actual speech recognition environment.

Stylization

Methods 1 and 2 use stylized versions of the F0 contour. The stylization works by selecting tonal turning points in the contour. The points are selected so that when reconnecting the points with straight lines, there may not, at any given point along the contour, be a difference in pitch between the reconstructed contour and the original contour that is larger than a set value, in this case one (1) semitone. This results in a series of time/frequency pairs, which describe the contour of a pitch pattern accurately, but with a smaller number of points than the full contour. See also Chapter 6.

8.4.3 Modelling

In order to build a classification model, we used the CART technique (described in Chapter 4) again. All in all, there were 1702 words. Of these we used 90% (1532 words) for training and saved 10% (170 words) for testing. As input we used both each parameter set individually and a combination of all parameters to see if there were synergy effects. Three runs where made, where we varied the 'stop' value[5], setting it to either 5, 10 or 20. The lower the value, the more fine-tuned to the training set the models get and there is a risk that the models get over-trained. We trained models to predict five different features:

[5] According to Black et al. (1998) this value "specifies the minimum number of examples necessary in the training set before a question is hypothesized to distinguish the group"

1. The village that the speaker of an utterance is from. There were 37 different villages in the material. Pure guessing thus has a 2.7% chance of being correct.

2. The province that the speaker of an utterance is from. There are 10 different provinces. Guessing has a 10% chance.

3. The accent type of a word. There are two different accents, Accent 1 or Accent 2. Guessing has a 50% chance.

4. The prominence level of a word. This is either focal or non-focal. Guessing has a 50% chance.

5. The dialect type, as classified in Gårding (1977) on the data by Meyer (1937, 1954). In the geographical region treated in the material (southern Sweden), there were four different categories. Chance is thus 25%.

The dialect type classification deserves some further comments. In Gårding's study, dialects were classified according to the timing and the number of peaks in Accent 2 words. We compared the location of the Swedia villages with the ones in Gårding's study. Sometimes, when villages were very close the choice was easy. In other cases, where none of Gårding's villages was on the exact location we had to choose a type by interpolating between the neighbouring villages. In a few cases we even had to examine our data briefly in order to make a selection. The results of the classification are presented in Appendix A.

8.4.4 Results

The results for each feature prediction are shown in Table 8.3. The individual results for each method and stop condition are shown in the cells of each part of the table. The results are presented as percentages of correct classifications. The best overall result for each predicted feature is printed in bold face.

8.4.5 Discussion

First, it should be noted that the best results for province and accent type are obtained when combining all the methods, whereas method 1 gets the best results for prediction of village, of prominence level and of Gårding's dialect types.

Village

The best result is 15.7%. This task is clearly a very difficult one, and much higher results should not be expected on the basis of any other parameterization. This task is difficult even for an expert human listener.

Table 8.3: Prediction results for Götaland material: village, province, word accent type, focus and Gårding's dialect types.

	Method	5	10	20
Village	All	13.5	12.4	9.7
	1	10.8	**15.7**	9.7
	2	10.3	10.8	11.9
	3	12.4	8.6	11.4
Province	All	**30.3**	23.2	28.7
	1	24.3	27	28.7
	2	23.8	23.2	20.6
	3	19.5	23.8	23.8
Word accent	All	78.2	78.2	**82.9**
	1	76.5	71.2	74.1
	2	65.3	70.0	65.9
	3	72.9	78.2	78.2
Focus	All	59.4	61.2	61.2
	1	62.4	67.1	**69.4**
	2	51.8	55.9	58.8
	3	54.1	56.5	52.9
Dialect type	All	55.2	55.9	61.2
	1	62.3	61.8	**65.3**
	2	57.1	62.3	59.4
	3	56.5	64.1	60.0

Province

The best result is 30.3%. Even though this is better than the estimated baseline result of 10% (since there are ten provinces, one tenth correct is roughly what we would get if we guessed on the same province in all trials) this is still quite poor. The task of predicting what province the speaker of a given utterance is from is probably too specific to be performed reliably on the basis of F0 of disyllabic words only.

Accent type

The best result is 82.9% correct. Compared with the 91.3% reached for unseen data from the five dialects in Section 8.3.2, we think this is rather good, given that the geographical spread is much higher, that more dialect types are represented in the material and that the prediction is based on prosodic information only, without geographical information.

Focus

The best result is 69.4% correct. The correct prominence level can thus be correctly predicted in 7 cases out of ten. This means it is somewhat more difficult to predict prominence level than to predict word accent.

Gårding's regional dialect types

The best result is 65.3%, which we also rate as rather good given that an estimated baseline result would give 25%. This task is less specific than predicting province and therefore easier. Note, however that Gårding's classification takes into account the number of peaks in a word, a parameter which is not used in this study.

8.5 Experiment IV: Classification of word accent, focus and dialect type, material from the whole Swedish language area

This study was similar to Experiment III, except that we used material from the whole Swedish-speaking language area. Note that this includes material from the Swedish-speaking areas in Finland, thereby adding a fifth type in the Gårding classification, meaning that the chance performance drops to 20%. In addition, the word accent distinction in these dialects is neutralized. This of course means that it is very hard to predict the word accent type in these dialects. In total, there were 4508 words, of which 90% (4058) were used in training and 10 % (450) for testing. As before, we varied the stop parameter, setting it to 5, 10, 15, 20 and 30. Furthermore, as we had more samples, an additional tree building technique was tested. When building trees, it is possible to first use a small stop value and build an over-trained tree. Then the tree is pruned (leaves are removed) so that it best matches a data set that is *held out*[6] from the original training data. This can produce better results as the stop value effectively is allowed to vary in different parts of the tree. We tested different variations of this, holding out 0%, 10% and 20% of the data. We did not pursue the tests of the features *village* and *province*, as this type of prediction clearly was too difficult and we only expected these results to be even lower.

For classification of Gårding's regional types, we also tested splitting up the data set into accented and focussed words. In effect, this means that we performed individual tests for Accent 1 words and Accent 2 words. Six such groups were created and split into training and testing sets. The reason that we did this is that we know that accent type and word accent vary along the same dimension, and that

[6]This follows the terminology used by Taylor et al. (1999) in describing decision tree building. It means that a separate subset of the training data is used for pruning.

one and the same intonation pattern may be of dialect type 1A if it is an Accent 1 word, but of type 2A if it is an Accent 2 word. Providing the classifier with this information should improve the performance. The following subgroups were thus also tested:

- U1 – Unfocussed accent 1 words.
- F1 – Focussed accent 1 words.
- U2 – Unfocussed accent 2 words.
- F2 – Focussed accent 2 words.
- All1 – All accent 1 words.
- All2 – All accent 2 words.

The same reasoning may be applied (reversely) to the word accent classification – one and the same intonation pattern may be an Accent 1 word if it is of type 1a, but an Accent 2 word if it is of type 2a. This created the following subgroups, which also were tested:

- 1A – Words in Gårding type 1A
- 2A – Words in Gårding type 2A
- 1B – Words in Gårding type 1B
- 2B – Words in Gårding type 2B
- 0 – Words in Gårding type 0

These latter subgroups were also tested for Focus prediction.

When the data is split in this way, the number of samples in each group becomes rather small. In order to perform a validity check of the results, tests were performed using several splits of the data. For word accent and focus we ran tests of 90%-10% and 50%-50% splits of the data. For Gårding's types we ran a 50%-50% split for the first four groups (U1, F1, U2 and F2) and a 90%-10% split of the others (All1 and All2), as there were more samples in this group. The number of samples in the test set is reported along with the percentage correct scores.

Table 8.4: Prediction results for the All-Swedish material: word accent type, focus and dialect types.

	Method	Result	Stop	Held out
Accent	All	**79.3**	15	10
	1	74.2	20	20
	2	64.7	30	20
	3	73.8	20	10
Focus	All	57.8	5	10
			10	20
	1	**62.2**	10	10
	2	59.8	30	10
	3	57.6	5	20
Dialect	All	58.4	20	20
type	1	**59.1**	30	10
	2	57.1	20	10
	3	57.1	20	10

8.5.1 Results

The results for the whole dataset are shown in Table 8.4. As there now are 15 different trees[7] for each method, we only present the best results for each. The best overall result for each predicted feature is printed in bold face.

The results for the individual datasets are shown in Table 8.5. We only present the best results for each sub-dataset. In the case of word accents, we only tested it using the combined method, as this performed much better than the individual methods for the whole dataset.

8.5.2 Discussion

Let us first state that the method of "holding out" data is very successful. Almost all the highest scores for all types of methods and predicted features are obtained when using this technique. Note that we did not use this technique in Experiment III, which means that the results are not completely comparable.

All-Swedish data

For word accent, the best result is obtained using a combination of all methods, the difference being more than five percentage units better. Method I perfoms better in the other cases. However, we think that the differences in these cases are quite small and we would be careful to rate Method I better than the others. The fact that one

[7] 5 stop conditions * 3 held out conditions = 15

Table 8.5: Prediction results for subsets of the All-Swedish material: word accent type, focus and dialect types.

	Group	Result	Size of test set	Method	Stop	Held out
Accent	1A	86.4	59	All	30	10
(90%-10%)	2A	80.6	217	All	30	10
	1B	85.4	41	All	10	0 and 10
	2B	80.5	118	All	20	20
	0	76.9	13	All	15,20 and 30	20
Accent	1A	82.3	299	All	10	10
(50%-50%)	2A	76.8	1086	All	15	0
	1B	80.4	208	All	10	20
	2B	78	591	All	15	0
	0	75.4	69	All	10	10
Focus	1A	62.7	59	Several	–	–
(90%-10%)	2A	67.7	217	Fall	30	20
	1B	78	41	All	15	20
	2B	66.9	118	Several	–	–
	0	76.9	13	Several	–	–
Focus	1A	60.2	299	2 (Point)	15	20
(50%-50%)	2A	62.2	1086	All	30	20
	1B	69.7	208	3 (Tilt)	10	20
	2B	65.3	591	1 (Fall)	20	20
	0	71	69	2 (Point)	10	20
Dialect	U1	55.8	611	1 (Fall)	10	10
type	F1	58.3	471	3 (Tilt)	20	10
	U2	53.7	681	1 (Fall)	20	10
	F2	59.4	490	1 (Fall)	30	10
	All1	56.9	216	2 (Point)	20	20
	All2	60.7	234	All	20	20

single method performs better than the combination suggests that the differences between the methods may be caused more than the nature of the data than their respective generalization capabilities.

The best result for Word Accent is 79.3% correctly classified words. This is obtained when using a combination of all the methods. This means that word accent can be predicted correctly in four out of five cases for this material, which contains words from the whole Swedish-speaking language area. Compared with the result for the geographically limited area (82.9%), there is only a slight performance decrease. We think that our methods are quite useful, or at least promising, for

classifying word accent category. As stated above (in Section 8.4.2), there might still be some room for improvement.

The best result for Focus is 62.2% correctly classified words. This is lower than the result for the geographically limited area in Experiment II. We think that this can be explained. In Experiment III, a larger proportion of the dialects are of type 1A and 1B, where focus is signalled as a modification of the pitch gesture for the word accent. When we use material from all of Sweden, we add more dialects of types 2A and 2B that signal focus by means of an extra gesture after the word accent. All of our methods concentrate on the pitch information in the stressed vowel where the pitch gesture for the word accent is usually found. This means that our methods may miss focus information that occurs outside of the word accent. Adding more dialects that signal focus in this way thus will lower the results.

For Gårding's dialect types our result is 59.1% correct, compared to 65.3% in Experiment III. The task in this experiment is more difficult, as there are more categories. The method can thus predict dialect type correctly in three cases out of five.

Subgroups

The prediction of word accent category is better for types 1A and 1B than for types 2A and 2B. This is true in both train/test distributions. The former, traditionally called "one-peaked" dialects score around 85%/81.5%, whereas the latter,"two-peaked", score around 80%/77.5%. It is thus somewhat easier to recognize word accent correctly in the "one-peaked" dialects. The results for the "one-peaked" groups are also better than for the All-Swedish material.

The Type 0 group scores lower, but still not as poorly as expected, considering that the accent difference normally is described as neutralized. However, studies by Selenius (1972) and Svärd (2001) has indicated that there might be small acoustic differences between the word accents in the dialects of Snappertuna and Närpes.

Focussed words are best distingushed from unfocussed words for dialect type 1b. There is some variation depending on the size of the test set among the other categories and there is no clear tendency of any word accent effect.

For dialect type, the Accent 2 group get a higher score than the Accent 1 group, and the groups with focussed words get higher scores than the groups without focus. It is thus more reliable to recognize dialect type using focussed words than nonfocussed. If the focus information is unavailable, Accent 2 words are more reliable than Accent 1 words.

8.6 Experiment V: Acoustic classification of word and focal accent using linear classifiers

Word accent and focal accent classification is a rather simple task: we only have a choice between two values: for word accent Accent 1 and Accent 2, and for focal accent whether it occurs or not. This means it is possible to use a simple linear classifier for any variable that is used. With a linear classifier it is possible to collect all the data for a given feature for a given village and then determine at which value the best split of the data occurs. It is then possible to compare the performances of all individual features. This analysis gives us the best performing feature for a given classification task for a given village.

8.6.1 Word accent

The results for word accent is shown in Table 8.6. For an explanation of the village abbreviations, and a map of Sweden and Finland with the locations of the villages, see Appendix A.

Table 8.6: Word accent classification. Best single feature, its performance and the number of occurrences of each category for each village. The "Acc" column means this accent is predicted if the actual value is lower than the threshold value.

Village	Best feature	Threshold	Acc	Correct	*n* A1	*n* A2	*n* tot
ank	t1_vs_vdur	0.0990	2	93.2%	22	28	50
anu	ft2_vs_vdur	1.8498	1	80.8%	23	18	41
ara	point1_st	2.1741	1	79.3%	17	21	38
are	ft2_vs_vdur	1.5700	1	85.7%	26	24	50
arj	ft2_vs_vdur	0.9513	1	93.8%	23	24	47
ars	tilt_dur	0.1550	2	92.3%	23	27	50
asb	fall_height	-7.5349	2	81.5%	13	27	40
asp	ft1_vs_vdur	-0.1526	1	92.3%	22	16	38
bar	tilt_dur	0.1400	2	86.3%	25	35	60
ben	ft1_vs_vdur	-0.2360	1	78.1%	19	26	45
ber	tilt_amp	12.8737	2	93.8%	18	16	34
bju	point1_st	-3.3937	1	87.2%	27	11	38
bod	tdur_vdur	0.7136	1	73.1%	15	14	29
bor	t2_vs_vdur	2.0072	2	81.3%	16	15	31
bra	tilt_dur	0.1650	2	83.3%	6	6	12

continued on next page

continued from previous page

Village	Best feature	Threshold	Acc	Correct	*n* A1	*n* A2	*n* tot
bre	tilt_dur	0.1350	1	90.6%	20	26	46
bro	ft1_vs_vdur	0.4331	1	88.8%	23	29	52
brr	tilt_tilt	-0.9583	2	80.7%	11	17	28
bur	ft2_vs_vdur	1.1627	1	93.5%	17	14	31
dal	ft1_vs_vdur	-0.2995	1	73.8%	65	59	124
del	ft1_vs_vdur	-0.0967	1	91.4%	21	24	45
dra	fall_start_st	1.2976	2	100.0%	1	2	3
fao	t1_vs	0.0198	1	92.1%	14	19	33
far	fall_start_st	4.2709	1	80.6%	23	24	47
fin	ft1_vs	-0.0328	1	79.7%	49	52	101
fja	ft2_vs_vdur	1.0592	2	81.7%	22	25	47
flo	tilt_tilt	-1.0000	2	80.0%	15	24	39
fol	tilt_dur	0.1250	1	85.7%	14	21	35
fra	t1_vs	0.0200	2	90.5%	14	17	31
fri	tilt_amp	15.7750	2	78.4%	28	34	62
fro	tilt_amp	13.0801	2	84.2%	19	19	38
gas	ft2_vs_vdur	0.7305	1	74.2%	43	46	89
gra	tilt_amp	16.9035	1	70.6%	20	18	38
grg	ft2_vs_vdur	2.5757	1	77.1%	27	20	47
grs	ft1_vs	-0.0288	2	78.4%	41	44	85
ham	tilt_dur	0.1300	1	89.2%	16	22	38
hao	ft1_vs_vdur	-0.1538	2	69.4%	46	43	89
har	ft1_vs	-0.0223	2	69.1%	52	69	121
hou	tilt_amp	10.1115	2	90.9%	11	7	18
hus	ttime_vs	0.0896	1	79.6%	35	34	69
ind	ft2_vs_vdur	1.3011	1	77.5%	16	10	26
jab	ft1_vs	-0.0361	1	64.9%	47	44	91
jam	ttime_vs	0.0847	1	97.7%	22	31	53
jar	ft1_vs_vdur	0.5630	1	83.7%	16	19	35
kaa	ft1_vs_vdur	-0.0177	1	86.6%	32	39	71
kar	tilt_dur	0.0900	2	76.1%	41	46	87
kol	ft2_vs_vdur	2.6644	2	83.5%	16	28	44
kor	t1_vs	0.0186	1	83.3%	21	27	48
kra	*	0	1	0.0%	6	0	6
kyr	fall_end_st	-6.9687	1	72.1%	13	16	29
lan	tilt_f0	100.7300	1	68.3%	32	31	63
lek	ttime_vs	0.0817	1	83.1%	39	34	73

continued on next page

continued from previous page

Village	Best feature	Threshold	Acc	Correct	*n* A1	*n* A2	*n* tot
lil	ft1_vs_vdur	-0.1343	1	92.9%	21	21	42
lod	ttime_vs_vdur	0.6024	2	93.2%	17	22	39
mal	fdur_vdur	1.2857	1	72.8%	61	65	126
nka	point2_st	0.1676	2	100.0%	8	1	9
nor	ft2_vs	0.2227	1	75.8%	48	33	81
nro	tilt_dur	0.1450	1	84.9%	30	34	64
nys	tdur_vdur	1.0382	2	82.1%	25	14	39
ock	tilt_dur	0.1350	2	82.0%	15	17	32
oru	ft1_vs_vdur	-0.1234	1	78.1%	20	16	36
oss	tilt_dur	0.1450	1	90.1%	18	23	41
ost	ft1_vs_vdur	-0.2471	1	87.5%	19	24	43
ova	ft1_vs_vdur	0.0428	1	88.0%	18	25	43
oxa	t1_vs	0.0192	2	83.7%	15	27	42
pit	stdur_vdur	0.4633	2	85.7%	29	7	36
rag	ft1_vs_vdur	-0.0430	2	83.3%	29	27	56
rim	tilt_amp	14.7045	2	78.8%	18	32	50
sal	tilt_amp	6.3298	1	100.0%	9	3	12
san	ft1_vs_vdur	-0.1353	2	83.3%	32	39	71
sar	tilt_amp	14.3445	2	97.2%	18	15	33
seg	tilt_dur	0.1250	2	86.2%	34	38	72
skb	fall_dur	0.0900	1	78.2%	20	21	41
ske	tilt_dur	0.1200	1	90.2%	30	37	67
sku	tilt_tilt	-0.3526	1	71.3%	59	56	115
sna	ft2_vs_vdur	2.2697	2	74.2%	14	19	33
soa	ft1_vs	-0.0238	2	71.6%	40	46	86
spr	tilt_dur	0.1300	1	84.2%	15	19	34
sta	ft1_vs_vdur	-0.1577	2	68.6%	27	25	52
ste	tdur_vdur	0.8376	1	84.3%	18	27	45
tja	ft1_vs_vdur	-0.0854	1	82.0%	18	25	43
toa	ft1_vs_vdur	0.0346	2	84.8%	13	23	36
toh	tilt_dur	0.1400	2	95.5%	13	22	35
too	tilt_dur	0.0900	2	77.1%	12	24	36
vac	tdur_vdur	0.9171	2	82.5%	19	26	45
vax	ttime_vs_vdur	0.5217	2	87.3%	26	34	60
vib	ft1_vs	-0.0244	2	78.7%	45	37	82
vil	ft1_vs	-0.0432	1	82.2%	43	39	82
vin	t1_vs_vdur	0.2729	2	94.4%	3	9	12

continued on next page

continued from previous page

Village	Best feature	Threshold	Acc	Correct	*n* A1	*n* A2	*n* tot
vvi	ft2_vs_vdur	1.0732	2	83.3%	24	30	54

As we can see, the best feature is different from village to village. Many of the features are used, mostly from methods 1 (Fall) and 3 (Tilt). The 'top' feature is ft1_vs_vdur (17 villages), followed by tilt_dur (15 villages). The average performance is 82.5%.

With this analysis, it is also possible to examine the performance of an individual feature for the whole set of villages. This is achieved by allowing different thresholds for each village for the features. This will give us information about what single feature is the best to use to describe the word accent difference for all villages. This is shown in Table 8.7. This analysis gives us the single best performing feature – given different thresholds – for the whole set of villages when the task is to distinguish between the word accents.

Table 8.7: Percentage of correctly classified word accents for all villages for each feature.

Feature	Correct
ft1_vs	75.1%
ft1_vs_vdur	74.5%
tilt_dur	71.1%
ft2_vs_vdur	68.7%
t1_vs	68.4%
tdur_vdur	68.4%
t1_vs_vdur	67.8%
tilt_amp	67.7%
ttime_vs	67.7%
tilt_tilt	67.5%
ft2_vs	66.9%
ttime_vs_vdur	66.8%
fdur_vdur	66.1%
sthei_stdur_vdur	65.2%
point_height	65.2%
stdur_vdur	64.8%
tilt_f0	64.8%
point0_st	64.8%

continued on next page

continued from previous page	
Feature	Correct
`fhei_fdur_vdur`	64.6%
`fall_start_st`	64.5%
`point1_st`	64.2%
`fall_height`	64.1%
`fall_dur`	64.1%
`t2_vs_vdur`	64.1%
`point_dur`	63.5%
`t2_vs`	63.1%
`point2_st`	62.9%
`fall_end_st`	62.2%

In Table 8.7, we see that the feature `ft1_vs` alone would classify word accent correctly in 75.1% cases. The feature `ft1_vs_vdur` performs almost as well, 74.5%. These two features are basically the same – the temporal location of the start of the largest fall in the word relative to the vowel onset – with the difference that `ft1_vs_vdur` is relative to the duration of the vowel. The feature `ft1_vs` represents, in effect, the location of the peak of a word accent.

8.6.2 Focal accent

The results for classifying focal/nonfocal word are shown in Table 8.8. Again, we see that the spread among the features is very large. Actually, 26 different features (of 28) are used. The average performance is 82.5%.

Table 8.8: Focus/no focus classification. Best single feature, its performance and the number of occurrences of each category for each village. The "Foc" column indicates what is predicted (0 = no focus, 1 = focus) if the actual value is lower than the threshold value.

Village	Best feature	Threshold	Foc	Correct	*n* 0	*n* 1	*n* tot
ank	`fdur_vdur`	1.2143	0	78.8%	31	19	50
anu	`fall_dur`	0.1200	1	74.4%	22	19	41
ara	`fhei_fdur_vdur`	-7.0030	0	70.5%	22	16	38
are	`fall_height`	-5.4672	0	72.5%	30	20	50
arj	`fall_end_st`	-2.0686	0	72.3%	29	18	47
ars	`fall_start_st`	1.9870	0	70.6%	21	29	50

continued on next page

continued from previous page

Village	Best feature	Threshold	Foc	Correct	*n* 0	*n* 1	*n* tot
asb	fdur_vdur	1.4583	0	83.0%	26	14	40
asp	point1_st	-1.0828	0	71.7%	24	14	38
bar	t2_vs	0.0784	1	71.0%	26	34	60
ben	fall_height	-6.8143	0	78.0%	25	20	45
ber	point2_st	-0.6333	0	73.3%	25	9	34
bju	fall_start_st	-0.1366	0	72.1%	26	12	38
bod	point_height	-2.6634	0	72.6%	15	14	29
bor	ft1_vs_vdur	-0.0316	1	84.2%	15	16	31
bra	ft2_vs_vdur	2.6869	1	87.5%	8	4	12
bre	point2_st	-2.8305	1	77.0%	34	12	46
bro	tilt_amp	15.8143	1	75.9%	32	20	52
brr	stdur_vdur	0.5880	0	72.3%	13	15	28
bur	point2_st	-1.9069	0	77.9%	16	15	31
dal	ttime_vs_vdur	0.7499	1	65.3%	59	65	124
del	sthei_stdur_vdur	-2.0308	0	72.0%	29	16	45
dra	fall_start_st	1.0228	1	100.0%	1	2	3
fao	fhei_fdur_vdur	-3.4520	0	88.6%	22	11	33
far	tdur_vdur	0.9530	0	74.3%	28	19	47
fin	tdur_vdur	0.7309	0	69.8%	48	53	101
fja	ttime_vs	0.0664	0	74.6%	31	16	47
flo	ft1_vs	0.1656	0	82.2%	23	16	39
fol	ft1_vs_vdur	0.1960	0	89.6%	24	11	35
fra	fdur_vdur	1.8095	0	75.9%	20	11	31
fri	tilt_f0	101.6280	0	66.1%	36	26	62
fro	ft2_vs	0.2270	0	71.1%	19	19	38
gas	point2_st	2.3399	0	81.5%	50	39	89
gra	ttime_vs	0.0250	0	70.3%	17	21	38
grg	fhei_fdur_vdur	-3.2457	1	69.8%	22	25	47
grs	point2_st	-2.7006	0	74.1%	50	35	85
ham	t2_vs	0.1101	0	72.5%	23	15	38
hao	fdur_vdur	1.5542	0	65.9%	53	36	89
har	tilt_dur	0.1000	0	72.1%	66	55	121
hou	ttime_vs_vdur	0.2048	1	87.5%	12	6	18
hus	tilt_tilt	0.2073	0	78.8%	41	28	69
ind	fall_height	-6.2557	1	71.3%	16	10	26
jab	t2_vs_vdur	0.8938	0	77.5%	48	43	91
jam	fall_height	-3.6386	1	75.6%	38	15	53

continued on next page

continued from previous page

Village	Best feature	Threshold	Foc	Correct	*n* 0	*n* 1	*n* tot
jar	tilt_amp	8.4100	1	74.1%	22	13	35
kaa	point0_st	6.2179	0	68.5%	43	28	71
kar	ft2_vs	0.3232	0	66.8%	47	40	87
kol	ft2_vs_vdur	2.5988	1	74.4%	15	29	44
kor	fdur_vdur	1.5833	1	76.4%	31	17	48
kra	fall_end_st	-0.8297	0	100.0%	4	2	6
kyr	point1_st	-0.7561	1	76.2%	14	15	29
lan	tilt_tilt	0.2811	0	73.0%	32	31	63
lek	ttime_vs_vdur	0.1640	0	71.1%	38	35	73
lil	fall_dur	0.1600	0	70.4%	25	17	42
lod	tilt_f0	108.1400	0	80.8%	24	15	39
mal	tilt_tilt	-0.0337	0	67.2%	62	64	126
nka	tilt_amp	6.7392	1	100.0%	6	3	9
nor	t2_vs	0.0898	0	65.3%	45	36	81
nro	fall_dur	0.1400	0	70.9%	34	30	64
nys	fall_start_st	2.1184	0	69.2%	26	13	39
ock	tilt_tilt	0.4768	0	77.4%	18	14	32
oru	ft2_vs	0.1572	1	66.7%	21	15	36
oss	stdur_vdur	0.7357	0	79.3%	29	12	41
ost	fdur_vdur	2.3000	0	93.1%	29	14	43
ova	point1_st	-0.0898	0	77.0%	27	16	43
oxa	fall_dur	0.2000	0	73.2%	28	14	42
pit	stdur_vdur	1.1628	0	83.1%	23	13	36
rag	fall_dur	0.1500	0	66.8%	39	17	56
rim	fall_start_st	2.7442	0	72.3%	31	19	50
sal	t1_vs_vdur	0.3751	1	92.9%	7	5	12
san	point2_st	-0.9699	0	70.5%	36	35	71
sar	fhei_fdur_vdur	-3.7584	1	84.6%	17	16	33
seg	t1_vs_vdur	0.1242	0	67.0%	45	27	72
skb	tilt_f0	87.3316	1	85.6%	15	26	41
ske	fall_dur	0.2400	0	73.1%	40	27	67
sku	tilt_dur	0.1300	0	69.3%	60	55	115
sna	tdur_vdur	0.8430	0	75.0%	22	11	33
soa	point1_st	4.9671	0	68.6%	48	38	86
spr	point2_st	2.4979	0	83.0%	22	12	34
sta	ft1_vs_vdur	0.5897	0	76.2%	28	24	52
ste	ft2_vs_vdur	1.3403	0	80.0%	30	15	45

continued on next page

continued from previous page

Village	Best feature	Threshold	Foc	Correct	*n* 0	*n* 1	*n* tot
tja	ft1_vs_vdur	-0.0815	0	75.2%	28	15	43
toa	point1_st	4.7303	0	74.7%	22	14	36
toh	tilt_tilt	-0.3888	1	71.0%	25	10	35
too	point0_st	7.8912	0	80.0%	25	11	36
vac	ft2_vs_vdur	1.2832	0	76.9%	27	18	45
vax	tdur_vdur	1.0798	0	63.3%	34	26	60
vib	tilt_amp	5.0103	0	69.9%	42	40	82
vil	ft1_vs_vdur	0.2911	0	71.1%	47	35	82
vin	tilt_f0	103.7740	0	88.9%	9	3	12
vvi	tilt_dur	0.1400	1	72.2%	27	27	54

The performance of each individual feature for the whole set of villages, given different threshold values for each village, is shown in Table 8.7. The feature with the best performance is point2_st. This corresponds to the F0 level (in semitones) of the second stylization point after the vowel onset. We could speculate that this is related to the fact that the second stylization point always is measured rather late in the word and that the focus gesture in many dialects is realized as a second gesture that comes after the word accent, temporally. However, as many of the features perform with the same accuracy, we are uncertain that this actual ordering of the features is of much significance.

Table 8.9: Percentage of correctly classified focus/no focus for all villages for each feature.

Feature	Correct
point2_st	65.1%
tilt_f0	64.8%
ft1_vs_vdur	64.4%
tilt_tilt	64.4%
ft1_vs	64.3%
fall_start_st	63.9%
ttime_vs_vdur	63.8%
fall_dur	63.7%
tilt_amp	63.7%
point1_st	63.7%
point0_st	63.7%

continued on next page

continued from previous page

Feature	Correct
ttime_vs	63.6%
ft2_vs	63.6%
point_height	63.4%
fall_end_st	63.3%
sthei_stdur_vdur	63.2%
fdur_vdur	63.1%
ft2_vs_vdur	63.0%
fhei_fdur_vdur	63.0%
t2_vs	63.0%
tdur_vdur	62.9%
point_dur	62.9%
fall_height	62.8%
stdur_vdur	62.7%
t2_vs_vdur	62.7%
tilt_dur	62.6%
t1_vs	61.7%
t1_vs_vdur	61.5%

8.6.3 Summary

The first experiment has shed some new light on the compound accent pattern in Southern Swedish. In the dialects furthest to the south, Accent 1 dominates in certain types of compounds, whereas in the most northern of the dialects examined (Broby), the same compounds get Accent 2. We then developed a method to distinguish acoustically between the Accent 1 and Accent 2 compounds in Experiment II. This method was based on stylizations of F0 patterns. Similar methods were then used to build classifiers for word accent, focal accent and dialect type. In Experiment III, we used utterances from more than 100 different speakers (all male) from 10 provinces in southern Sweden. The utterances were focal and non-focal Accent 1 and Accent 2 words. The best results for village, province, accent type, focus and Gårding type were 15.7%, 30.3%, 82.9%, 69.4% and 65.3% correct, respectively. This showed that the categories of province and village are too specific to use for a prediction model based on F0, whereas accent type is possible to predict with rather high accuracy. In a more extended study (Experiment IV), we used utterances from around 250 speakers (all male) from the whole Swedish-speaking area. The best prediction results were: 79.3% for word accent, 62.2% for focus and 55.9% for Gårding's dialect types. Splitting up the data showed that accent prediction is somewhat more reliable for types 1A and 1B, the "one-peaked" dialects. Focus is best predicted in dialects of type 1B, and dialect type gets a higher prediction rate

correct for focussed words and for Accent 2 words. Finally, in Experiment V we analysed the performance of each feature individually for each village, in order to find the best feature for distinguishing A1 from A2. It was discovered that the best single feature is `ft1_vs`, which corresponds to the beginning of a fall in the stressed vowel.

The results presented here are valid for elicited, semi-spontaneous speech. Lexical variation is very limited, in fact only one example word of each word accent is used. Speaker variation, on the other hand, is very large. We have more than 250 different speakers (all male). An interesting topic for future studies would be to test these methods on longer stretches of spontaneous speech. Another issue would be to compare the result of this automatic classifier with the performance of humans.

Chapter 9

Concluding summary

In this thesis we have investigated methods for the computational modelling of Swedish prosody with speech technology in mind. We will now conclude with a summary of the major findings and contributions of this work. The summary follows the structure of the thesis by separating the issues of lexical and acoustic prosody modelling.

9.1 Lexical modelling of prosody

Lexical modelling of prosody has been investigated with perspectives from text processing, from phonology, and from letter-to-sound conversion.

Chapter 2 discusses the prediction of word prosody, in a text-to-speech context, for text words with a 'non-standard' orthography. These words include all uses of non-alphanumerical characters, such as %, -, : etc, and, in general, exhibit a mix of alphabetical, numerical and non-alphanumerical characters. These words pose some peculiar problems for pronunciation in general and also need some special treatment with regard to prosody. It is also argued that viewing the text processing phase as a text 'normalization' phase whose task is to convert 'non-standard' into alphabetic forms is problematic in a text-to-speech context and will only postpone, and even render more difficult, the assignment of proper prosodic structure. The chapter is concluded with a small investigation of the modelling of prosody in multi-compound words in some existing text-to-speech synthesis systems.

In Chapter 3, an analysis of the phonology of Swedish stress is presented. We have used a modern phonological theory, optimality theory, where constraints are ranked in order to evaluate several candidates which are in dynamic competition. The outcome of an optimality theoretic analysis can be said to be more dependent on the candidates' relations to each other than on their intrinsic features. The study departs from metrical phonology. The following basic metrical facts of Swedish are used: Swedish has bounded feet (the fundamental pattern is bisyllabic), is quantity sensitive (interacts with the weight of the rhyme of a syllable), the foot head is

left-bounded (trochaic), the word-head is right-bounded, and the foot-forming starts at the right edge of the word. This metrical analysis is then extended with weight analysis, using mora structure and the open/closed status of the syllable in order to determine syllable weight. Therefore, we avoid the 'chicken-and-egg' situation that arises from using vowel quality to determine stress position. The analysis shows that Swedish avoids final stress in monomorphemic words, except when a superheavy syllable occurs in final position. Further, it is noted that a heavy penultimate syllable often causes stress on the penult. With an open penult, the pattern is more complicated and we must rely on prespecified lexical information or other factors, such as submorphemic patterns or schwa syllables. Compound words, which are left-headed in Swedish, are treated on the clitical group level. We also briefly consider the structure of derivatives, which affect the stress pattern in different ways, with the affix either attracting or repelling stress, and possibly causing a compound stress pattern. All these facts are treated within an optimality theoretic framework, and the system is couched in constraints. We also present a ranking hierarchy that accounts for the stress pattern of Swedish.

The final chapter (Chapter 4) on the lexical modelling of prosody is also the largest. In this chapter, we developed an automatically trained rule-based system that predicts the lexical-prosodic properties of vowel length, stress position and word accent category, as well as allophonic segments, from a graphemic representation. Now, the prediction of vowel quality and consonant allophones obviously fall outside of prosody proper and therefore, apparently, also outside of the overall scope of this thesis. However, in this case, we argue that the prediction of prosody is rather entangled with the prediction of allophones. The results of Black et al. (1998) indicate that a combined model, where segments and prosody are predicted simultaneously, performs better than separate models for allophones and prosody that are applied sequentially. Furthermore, for the purpose of text-to-speech conversion, a complete system that predicts both allophones and prosody is clearly of more use than a system that only predicts prosody.

The model is built from a computer-readable pronunciation dictionary by using a machine learning technique called Classification and Regression Trees (CART). This method examines a large number of letter-to-sound mappings derived from the dictionary and uses the contexts of the letters in order to produce a hierarchically ordered set of rules. For allophones, prediction is done through a letter-by-letter method, where each letter predicts an allophone and the result is obtained through the concatenation of all individual predictions.

The results are 96.87% correctly predicted allophones, and 72.26% correctly predicted words. In the latter case the requirement is that all allophones in the word must be correct. For prosody, the scores on the word level are: 73.56% for stress patterns, 68.94% for word accent + stress pattern, and 55.24 for full allophone+prosody prediction. Prediction based on whole-word patterns performs

better for prosody. Position of main stress is correctly predicted in 88.6% of the cases, whereas full stress pattern gets a conjectured 'not-worse-than' score of around 84% and word accent+stress gets no worse than 72%.

9.2 Acoustic modelling of prosody

Our contribution to the acoustic modelling of prosody includes one theoretical and one empirical section. The theoretical part consists of two surveys, one on previous work on speech-technology related modelling of Swedish intonation, and one on the major intonation models in the field.

The review of Swedish intonation models presents the work by Carlson and Granström (1973), Bruce (1977) and its later developments, Lyberg (1981) as well as the attempts at superpositional modelling by, e.g., Gårding (1983), Ljungqvist and Fujisaki (1993) and Fant et al. (2002). The work following Bruce (1977) is most extensively described and starts off with the original characterization of the pitch rules, the prominence levels and the relation to syllabic elements. This is followed by the introduction of downstepping and phrasing elements, elaboration of the modelling of focus accent and a reformulation of the model in terms of autosegmental phonology. Finally, the work on model-based resynthesis that leads to rules for the distribution of tonal elements and their temporal and F0 level properties is described. Furthermore, work on reference tracking for phrase-level focus prediction as well as higher-level prosodic categories by, among others, Horne and Filipsson (1994, 1996) is treated.

Our own contribution within this area is the INTRA transcription and labelling environment and the IDL, a system designed for the specification of relations between abstract phonological labels and their realizations as F0 patterns. INTRA allows segmental-phonetic and tonal transcriptions to be fed into a speech synthesis system, where the phonetic labels control the temporal and segmental contents, and the tonal labels are used to produce an F0 contour through the intonation model. Subsequently, the resulting speech signal is a result of the provided transcriptions. The use of resynthesis from both the phonetic and the tonal transcriptions yields a fast estimation of how accurate the transcription is, and the result of a minor modification may be obtained almost instantly. The possibility of combining the labelling and the evaluation environments provides an efficient platform for speech transcription. Furthermore, the creation of prosodically varying speech stimuli for perception experiments is greatly simplified.

The intonation model was implemented in the IDL, so that it would be possible to generate different intonational characteristics of various dialects of Swedish. This is done by having a dialect-specific rule set for the mapping of the phonological transcription into F0 events. Some of the major dialects of Swedish are included in the F0 generation scheme, and thus we may simulate the intonation of different

dialects. The result is that we are able to produce synthesized utterances exhibiting dialectal variation while preserving the same phonological transcription. This also shows how close we can get to producing dialectal variation by varying intonation only.

In Chapter 6, we reviewed some of the most significant theories of acoustic intonation modelling. This survey includes the IPO model, the superpositional model of Fujisaki-Öhman, ToBI, INTSINT-MOMEL, Tilt and contour-faithful stylization models.

The IPO model is based on a perceptual analysis of intonation. Pitch contours are simplified by replacing portions of them with straight lines, while fulfilling the demand that they should remain perceptually identical to the original contour. An intonation grammar, where pitch movements are linked to linguistic functions, may then be specified. ToBI is basically a phonologically grounded model, that relates a prosodic transcription system to intonational properties. The development of ToBI is triggered by the need to label prosody in a standardized way. ToBi describes tonal categories using two pitch levels, High and Low, and further notational devices specify how these elements are linked to stressed syllables and phrase-level phenomena such as downstep. Prosodic coherence is annotated using numerical indices. Several of the models (Fujisaki-Öhman, Tilt and INTSINT-MOMEL) parameterizes the shapes of the pitch accent-related intonation events that occur in speech. This is done by measuring the amplitudes and the temporal properties of these events. Generation of pitch contours is then achieved by feeding these parameters into mathematical functions that generate smooth curves. In addition, the superpositional model presupposes a baseline contour on which local pitch modifications are overlayed. In contour-faithful stylization models, the F0 pattern is captured by extracting temporal and frequency information that originates from actual F0 contours. Two methods in this category were examined: the model by D'Alessandro and Mertens (1995) and the model implemented in the PRAAT computer program (Boersma and Weenink, 2003). The first one of these stylizes the F0 contour by the continuous addition of pitch points from the original contour at the temporal locations where the original and the on-going stylization contours differ most. The second one instead removes points where such a removal causes the least change in the pitch curve. In both methods, modifications end when the difference passes a given threshold. A comparison showed that the two stylization methods, in practice, yield identical results.

At the end of this survey, we select models for use in the studies of the coming chapters using the two criteria that they should not have been tried on Swedish before, and that they should provide an automatic analysis method. Two models fulfilled these requirements: the stylization model and the tilt model. These were consequently chosen for the continued studies.

Chapter 7 presents a study on the modelling of intonation for speech synthesis

for Swedish. The chapter describes an initial attempt at the construction of a data-driven model of Swedish intonation. The study is mainly concerned with model-building and prediction of the intonation patterns of accented words in a corpus of read news in Swedish. Extraction of pitch information is achieved by performing a stylization of the pitch contours. The information is used to build a model for the prediction of pitch patterns using linguistic features such as accent type and position of stress. The model is tested against unseen data from the same corpus. The evaluation is done by numerical comparisons. The RMSE between predicted and original contours for the different categories ranges between 3.7 and 31.4 Hz. The results show a clear tendency that the lower the order of stylization, the lower the RMSE. For word that are stylized with three or fewer points, the RMSE is below 10 Hz and 1 semitone.

In Chapter 8, directed towards recognition of prosody, we progress from an investigation of the realization of word accents in complex words in material from southern Sweden to the development of an intonation-based recognition system for prosodic and dialectal categories for material from the whole Swedish-speaking area in Sweden and Finland. This chapter consists of a collection of investigations directed towards the model-based description and prediction of prosodic variation in Swedish. All the material used in this chapter is taken from the Swedia 2000 dialect project.

The first investigation tries to clarify the status of compound accentuation in southern Swedish. This region is special in that word Accent 1 is used on certain complex (compound) words, that in other dialects have Accent 2. Material from five villages in southern Sweden was examined. The study indicates that Accent 1 is prevalent in three of them (Bara, Löderup and Norra Rörum), while it is virtually nonexistent in a fourth one (Broby). In the last village (Våxtorp) the usage is unstable. Different speakers use different accent patterns.

In the second study, the acoustic properties of the difference between the word accents from the previous study are formalized. This leads to a system for the prediction of word accent categories from intonational data. Using the stylization method developed in Chapters 6 and 7, a rule-based system is developed that predicts the correct word accent category in 91.3% for unseen data from southern Sweden. Furthermore, the idea behind linear classifiers for the characterization of intonational data is illustrated.

Material from a larger region in Sweden, namely the Götaland region, is examined in the next experiment. As we now have a dialectally more varied material, it also becomes interesting to examine the predictability of dialect types. Using three different parameterization methods, we develop prediction systems for regional (village, province and the larger-scale regions based on Gårding (1977)) and word-prosodic (word accent and focussed/unfocussed distinction) categories. The best results for village, province and Gårding type were 15.7%, 30.3% and 65.3%

correct predictions, respectively. Even though this is better than chance, this shows that the categories of village and province are too specific for a prediction model based on these F0 parameters, whereas the larger-scale regions represented by the Gårding types, though still only correct in about two-thirds of the cases, are easier to predict correctly. Word accent distinction is correctly predicted in 82.0% of the cases and prominence level in 69.4% in material from this region.

The next experiment was rather similar, but now we used material from the whole Swedish-speaking area. This material included more than 250 different speakers from almost one hundred different villages in Sweden and Finland. We also discontinued examining the predictability of the *village* and *province* categories, as the previous study showed that these were too specific to predict accurately. The best results for word accent, prominence level and Gårding type were 79.3%, 62.2% and 59.1% correct predictions, respectively. By examining subgroups of each category, it was also found that word accent was somewhat easier to predict in the "one-peaked" dialect groups (1a and 1b, the "SOUTH" and "DALA" groups), and that focussed/unfocussed was easier in the 1b ("DALA") group. For dialect type, better results were obtained when using focussed words only than when using unfocussed words, and, similarly, Accent 2 words were classified correctly more often than Accent 1 words. These results show the accuracy with which word accent, focus level and large-scale dialect type may be predicted by using intonational parameters derived from stylization models and the Tilt model. Word accent is thus the easiest to determine with a purely intonation-based method. As the parameterization methods used concentrate on the stressed syllable, we think that the methods used are well-suited for detecting word accent. Optimization of the method regarding the identification of the most suitable stylization point is a possible way of improvement. Focus level is somewhat accurately predicted, but here other possible correlates have been suggested, such as intensity, duration and spectral emphasis. This may be included in a focus detector to help to increase the performance. Dialect type, on the whole, is rather difficult, but as dialects also show a large variation in their segmental properties there is much additional phonetic information that could be used for dialect detection.

In the final study, linear classifiers were re-examined in order to find the best measurement feature for each village for distinguishing between the word accents and the prominence levels. For word accents, we found several villages where one single feature was enough to correctly describe more than 90% of the data within that village. The single best feature was found to be the temporal location of the start of the largest F0 fall in the word relative to the vowel onset – in essence, the timing of a peak in the F0 contour in a stressed syllable. For prominence levels, a similar analysis showed less accurate results per village, as well as a less clear indication of what the best single feature is. This is probably related to the fact that focus realisation is done in different ways in the dialects of Swedish.

Appendix A

Villages in Swedia

Table A.1: Villages and their classifications. Alternative classifications are in parentheses. This indicates that the classification was based on examination of data.

Number	Key	Village Name	Province	Type
1	ank	Ankarsrum	Småland	2A
2	anu	Anundsjö	Ångermanland	2A
3	are	Åre	Jämtland	2A
4	arj	Arjeplog	Lappland	2A
5	ars	Årstad-Heberg	Halland	1A (2A)
6	ara	Årsunda	Gästrikland	2A
7	asb	Asby	Östergötland	2B
8	asp	Aspås	Jämtland	2A (1A)
9	bar	Bara	Skåne	1A
10	ben	Bengtsfors	Dalsland	2B
11	ber	Berg	Jämtland	2B
12	bju	Bjurholm	Ångermanland	2A
13	bod	Böda	Öland	2A
14	bor	Borgå	Nyland	0
15	bra	Brändö	Åland	0
16	bre	Bredsätra	Öland	2A
17	bro	Broby	Skåne	1A
18	brr	Burseryd	Småland	1A (2B)
19	bur	Burträsk	Västerbotten	2A (1B)
20	dal	Dalby	Värmland	2A
21	del	Delsbo	Hälsingland	2A
22	dra	Dragsfjärd	Åboland	0

continued on next page

continued from previous page

Number	Key	Village Name	Province	Type
23	far	Färila	Hälsingland	2A
24	fao	Fårö	Gotland	1B
25	fja	Fjällsjö	Ångermanland	2A
26	flo	Floby	Västergötland	2B
27	fol	Fole	Gotland	1B
28	fra	Frändefors	Dalsland	2B
29	fri	Frillesås	Halland	2B
30	fro	Frostviken	Jämtland	2A
31	gas	Gåsborn	Värmland	2B
32	grg	Grangärde	Dalarna	1B
33	gra	Gräsmark	Värmland	2B
34	grs	Gräsö	Uppland	2A
35	hao	Hammarö	Värmland	2B
36	ham	Hamneda	Småland	1A (2B)
37	har	Haraker	Västmanland	2A (1B)
38	hou	Houtskär	Åboland	0
39	hus	Husby	Dalarna	1B
40	ind	Indal	Medelpad	2A
41	jam	Jämshög	Blekinge	1A
42	jab	Järnboås	Västmanland	2B
43	jar	Järsnäs	Småland	1A (2B)
44	kaa	Kärna	Bohuslän	2B
45	kar	Kårsta	Uppland	2A
46	kol	Köla	Värmland	2B
47	kor	Korsberga	Västergötland	2B
48	kra	Kramfors	Ångermanland	2A
49	kyr	Kyrkslätt	Nyland	0
50	lan	Länna	Södermanland	2A
51	lek	Leksand	Dalarna	1B
52	lil	Lillhärdal	Härjedalen	2A
53	lod	Löderup	Skåne	1A
54	mal	Malung	Dalarna	1B
55	mun	Munsala	Österbotten	0
56	nap	Närpes	Österbotten	0
57	nka	Nederkalix	Norrbotten	2B
58	nlu	Nederluleå	Norrbotten	2A
59	nor	Nora	Uppland	2A

continued on next page

continued from previous page

Number	Key	Village Name	Province	Type
60	nro	Norra_Rörum	Skåne	1A
61	nys	Nysätra	Västerbotten	2A
62	ock	Ockelbo	Gästrikland	2A
63	oru	Orust	Bohuslän	2B
64	oss	Össjö	Skåne	1A
65	ost	Östad	Västergötland	2B
66	ova	Ovanåker	Hälsingland	2A
67	oka	Överkalix	Norrbotten	0
68	oxa	Öxabäck	Västergötland	2B
69	pit	Piteå	Norrbotten	2A
70	rag	Ragunda	Jämtland	2B
71	rim	Rimforsa	Östergötland	2B
72	sar	Särna	Dalarna	2A
73	seg	Segerstad	Öland	2A
74	ske	Skee	Bohuslän	2B
75	skb	Skinnskatteberg	Västmanland	2A (1B)
76	sko	Skog	Hälsingland	2A
77	sku	Skuttunge	Uppland	2A
78	sna	Snappertuna	Nyland	0
79	fin	Södra_Finnskoga	Värmland	2A
80	soa	Sorunda	Södermanland	2A
81	spr	Sproge	Gotland	1B
82	san	S:t_Anna	Östergötland	2A
83	ste	Stenberga	Småland	1A (2A)
84	sta	Stora_Mellösa	Närke	2A
85	sto	Storsjö	Härjedalen	2A
86	stm	Ström	Jämtland	2A
87	tja	Tjällmo	Östergötland	2B
88	toh	Torhamn	Blekinge	1A (2B)
89	tor	Torp	Medelpad	2A
90	toa	Torsås	Småland	2B
91	too	Torsö	Västergötland	2B
92	vac	Väckelsång	Småland	2B
93	vax	Växtorp	Halland	1A
94	vem	Vemdalen	Härjedalen	2A
95	vib	Viby	Närke	2A
96	vim	Vilhelmina	Lappland	2A

continued on next page

continued from previous page

Number	Key	Village Name	Province	Type
97	vil	Villberga	Uppland	2A
98	vin	Vindeln	Västerbotten	2A
99	vvi	Västra_Vingåker	Södermanland	2A
100	vor	Vörå	Österbotten	0

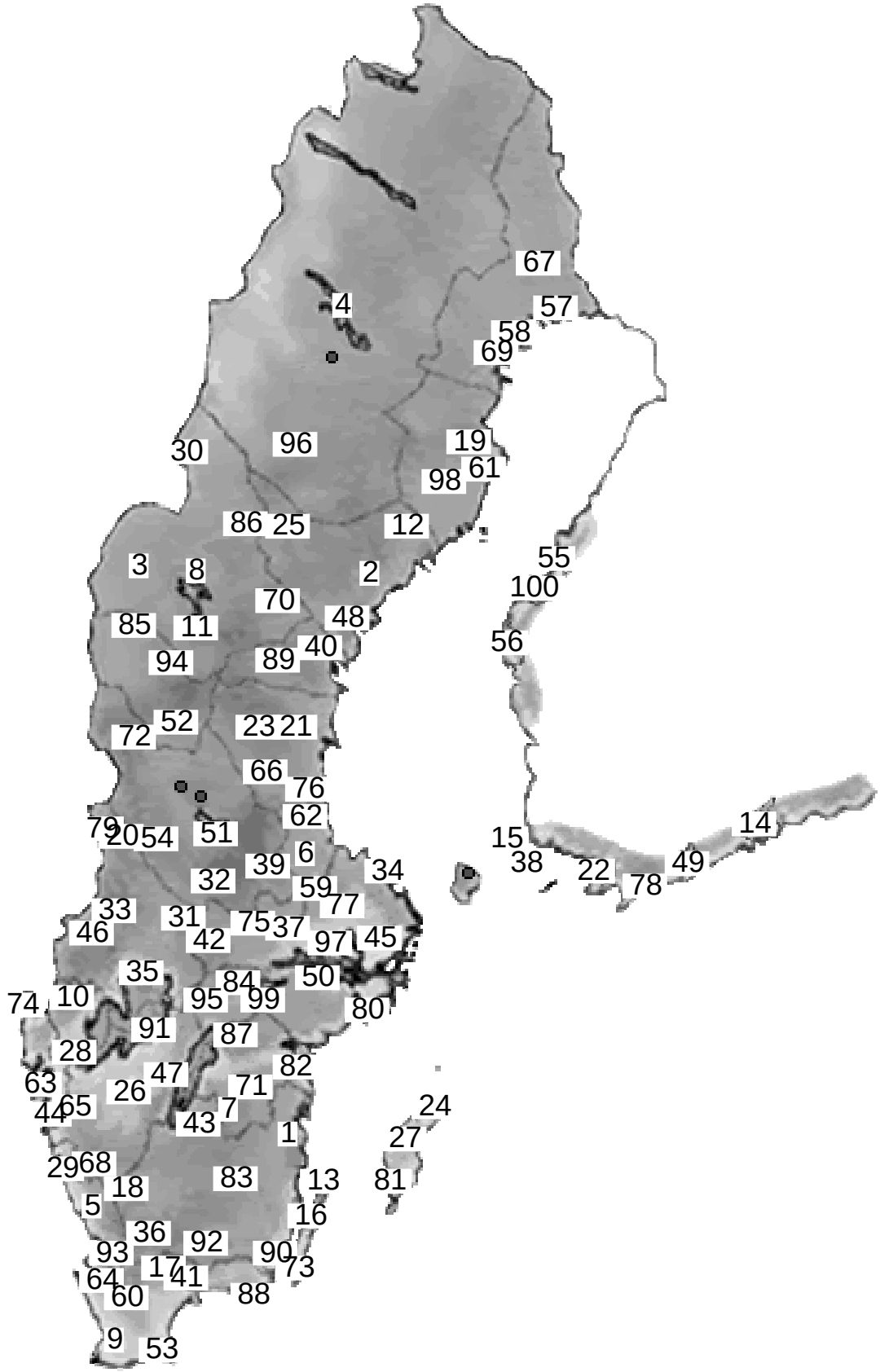

Figure A.1: Map of Sweden and Finland with the locations of the villages in the Swedia material.

Bibliography

Allen, J., Hunnicutt, M. S., and Klatt, D. (1987). *From text to speech: The MITalk system*. Cambridge University Press.

Allén, S. (1970). *Frequency Dictionary of Present-Day Swedish*. Almqvist & Wiksell, Stockholm.

Anward, J. and Linell, P. (1976). Om lexikaliserade fraser i svenskan. *Nusvenska studier*, 55–56:77–119.

Bagshaw, P. C., Hiller, S. M., and Jack, M. A. (1993). Enhanced pitch tracking and the processing of F0 contours for computer aided intonation teaching. In *Proceedings of Eurospeech 93*, pages 1003–1006, Berlin, Germany.

Bannert, R. (1998). Two thousand and one syllables in spoken standard Swedish: aspects of syllabification. *PHONUM*, (6):51–82.

Batliner, A., Möbius, B., Möhler, G., Schweitzer, A., and Nöth, E. (2001). Prosodic models, automatic speech understanding, and speech synthesis: towards the common ground. In *Proceedings of Eurospeech 2001*, volume 4, pages 2285–2288, Aalborg, Denmark.

Beaugendre, F. (1994). *Une étude perceptive de l'intonation du français*. PhD thesis, L'Universitè Paris XI.

Bernstein, J. and Pisoni, D. B. (1980). Unlimited text-to-speech system: Description and evaluation of a microprocessor-based device. In *Proceedings of ICASSP-80*, pages 576–579, Denver, CO.

Black, A. and Hunt, A. (1996). Generating F0 contours from ToBI labels using linear regression. In *Proceedings of ICSLP 96*, volume 3, pages 1385–1388, Philadelphia.

Black, A. and Lenzo, K. (2003). Building synthetic voices. Website: http://www.festvox.org/festvox/index.html, 1999-2003.

Black, A., Lenzo, K., and Pagel, V. (1998). Issues in building general letter to sound rules. In *Proceedings of 3rd ESCA Workshop on Speech Synthesis*, pages 77–80, Jenolan Caves, Australia.

Boersma, P. (1993). Accurate short-term analysis of the fundamental frequency and the harmonics-to-noise ratio of a sampled sound. *Proceedings of the Institute of Phonetic Sciences University of Amsterdam*, 17:97–110.

Boersma, P. and Weenink, D. (2003). PRAAT: doing phonetics by computer. Website: http://www.praat.org, 1992-2003.

Breen, A., Eggleton, B., Dion, P., and Minnis, S. (2002). Refocussing on the text normalisation process in text-to-speech systems. In *Proceedings of ICSLP 02*, pages 153–156, Denver, USA.

Breiman, L., Friedman, J., Olshen, R., and Stone, C. (1984). *Classification and regression trees*. Wadsworth and Brooks.

Brodda, B. (1979). Något om de svenska ordens fonotax och morfotax: Iakttagelser med utgångspunkt från experiment med automatisk morfologisk analys. *PILUS*, (38).

Bruce, G. (1973). Tonal accent rules for compound stressed words in the Malmö dialect. In *Working Papers*, number 7, pages 1–35. Phonetics Laboratory, Lund University.

Bruce, G. (1974). Tonaccentregler för sammansatta ord i några sydsvenska stadsmål. In Platzack, C., editor, *Svenskans beskrivning*, number 8, pages 62–75.

Bruce, G. (1977). *Swedish Word Accents in Sentence Perspective*. CWK Gleerup.

Bruce, G. (1982). Developing the Swedish intonation model. In *Working Papers*, number 22, pages 51–116. Department of Linguistics, Lund University.

Bruce, G. (1983). Accentuation and timing in Swedish. *Folia Linguistica*, XVII(1–2):221–238.

Bruce, G. (1987). How floating is focal accent? In Gregersen, K. and Basbøll, H., editors, *Nordic Prosody IV*, pages 41–49. Odense University Press.

Bruce, G. (1993). On Swedish lexical stress patterns. *PHONUM*, 2:41–50.

Bruce, G. (1998). *Allmän och svensk prosodi*. Number 16 in Praktisk Lingvistik. Department of Linguistics and Phonetics, Lund University.

Bruce, G., Engstrand, O., and Eriksson, A. (1998). De svenska dialekternas fonetik och fonologi år 2000 (Swedia 2000) - en projektbeskrivning. In *Proceedings of 6:e Nordiska Dialektolog-konferensen*, pages 33–54.

Bruce, G., Filipsson, M., Frid, J., Granstrom, B., Gustafson, K., Horne, M., and House, D. (2000). Modelling of Swedish text and discourse intonation in a speech synthesis framework. In Botinis, A., editor, *Intonation. Analysis, Modelling and Technology*, pages 291–320. Kluwer Academic.

Bruce, G., Granstrom, B., Gustafson, K., Horne, M., House, D., and Touati, P. (1997). On the analysis of prosody in interaction. In Sagisaka, Y., Campbell, N., and Higuchi, N., editors, *Computing Prosody*, pages 43–59. Springer.

Bruce, G. and Granström, B. (1989). Modelling Swedish intonation in a text-to-speech system. *KTH, STL-QPSR*, 1:17–21.

Bruce, G. and Granström, B. (1990). Modelling Swedish prosody in text-to-speech: Phrasing. In Wiik, K. and Raimo, I., editors, *Nordic Prosody V*, pages 26–35. Phonetics Department, Turku University.

Bruce, G. and Granström, B. (1993). Prosodic modeling in Swedish speech synthesis. *Speech Communication*, 13:63–73.

Bruce, G., Granström, B., Filipsson, M., Gustafson, K., Horne, M., House, D., Lastow, B., and Touati, P. (1995). Speech synthesis in spoken dialogue research. In *Proceedings of Eurospeech 95*, pages 1169–1172, Madrid, Spain.

Bruce, G. and Gårding, E. (1978). A prosodic typology for Swedish dialects. In Gårding, E., Bruce, G., and Bannert, R., editors, *Nordic Prosody*, pages 219–228. Department of Linguistics, Lund University.

Campbell, N. and Venditti, J. (1995). J-ToBI: an intonational labeling system for Japanese. Paper presented at the Autumn meeting of the Acoustical Society of Japan.

Campione, E., Hirst, D., and Veronis, J. (2000). Automatic stylisation and modelling of french and italian intonation. In Botinis, A., editor, *Intonation. Analysis, Modelling and Technology*, pages 185–208. Kluwer Academic.

Carlson, R. and Granström, B. (1973). Word accent, emphatic stress, and syntax in a synthesis by rule scheme for Swedish. *KTH, STL-QPSR*, 2–3:31–36.

Carlson, R. and Granström, B. (1976). A text-to-speech system based entirely on rules. In *Conference record 1976 IEEE International Conference on ASSP*, pages 686–688, Philadelphia, PA, USA.

Carlson, R. and Granström, B. (1986). A search for durational rules in a real-speech data base. *Phonetica*, 43:140–154.

Carlson, R. and Granström, B. (1989). Modeling duration for different text materials. In *Proceedings of Eurospeech 89*, volume 2, pages 328–331, Paris, France.

Cohen, A. and Hart, J. 't. (1967). On the anatomy of intonation. *Lingua*, (19):177–192.

Coker, C. H., Church, K. W., and Liberman, M. Y. (1990). Morphology and Rhyming: Two Powerful Alternatives to Letter-to-Sound Rules for Speech Synthesis. In *Proceedings of the ESCA Workshop on Speech Synthesis*, pages 83–86, Autrans, France.

Daelemans, W. and Van den Bosch, A. (1996). Language-independent data-oriented grapheme-to-phoneme conversion. In van Santen, J. P. H., Sproat, R. W., Olive, J. P., and Hirschberg, J., editors, *Progress in Speech Synthesis*, pages 77–89. Springer Verlag.

Daelemans, W. and Van den Bosch, A. (2001). Treetalk: Memory-based word phonemisation. In Damper, R., editor, *Data-Driven Techniques in Speech Synthesis*, pages 149–172. Kluwer Academic Publishers.

D'Alessandro, C. and Mertens, P. (1995). Automatic pitch contour stylization using a model of tonal perception. *Computer, Speech and Language*, 9:257–288.

Damper, R. I., Marchand, Y., Anderson, M. J., and Gustafson, K. (1999). Evaluating the pronunciation component of text-to-speech systems for english: a performance comparison of different approaches. *Computer, Speech and Language*, 13(2):155–176.

Dusterhoff, K. (2000). *Synthesizing Fundamental Frequency Using Models Automatically Trained from Data*. PhD thesis, University of Edinburgh.

Dusterhoff, K., Black, A., and Taylor, P. (1999). Using decision trees within the tilt intonation model to predict F0 contours. In *Proceedings of Eurospeech 99*, pages 1627–1630, Budapest, Hungary.

Dutoit, T. (1997). *An introduction to text-to-speech synthesis*. Kluwer Academic, Dordrecht.

Eisfelder, R. and Hendrickson, T. (1999). The effect of prosody on listening comprehension of second grade students. Paper presented at the meeting of the ILLOWA Undergraduate Psychology Research Conference Website: http://homepages.culver.edu/illowa/abstr99.htm#Listening.

Ejerhed, E., Källgren, G., Wennstedt, O., and Åström, M. (1992). The linguistic annotation system of the Stockholm-Umeå corpus project. Report 33. Technical report, Department of Linguistics, Umeå University.

Eklund, R. and Lindström, A. (2001). Xenophones: An investigation of phone set expansion in Swedish and implications for speech recognition and speech synthesis. *Speech Communication*, 35(1–2):81–102.

Eklund, R. and Lyberg, B. (1995). Inclusion of a prosodic module in spoken language translation systems. In *The Journal of the Acoustical Society of America*, volume 98, pages 2894–2895.

Fant, G. and Kruckenberg, A. (2001a). F0 analysis and prediction in Swedish prose reading. In Grönnum, N. and Rischel, J., editors, *To Honour Eli Fischer-Jörgensen*, Travaux du Circle Linguistique de Copenhague, pages 124–147. Reitzel Copenhagen.

Fant, G. and Kruckenberg, A. (2001b). A novel system for F0 analysis and prediction. In *Proceedings of Fonetik 2001*, pages 38–41, Örenäs, Sweden.

Fant, G., Kruckenberg, A., Gustafson, K., and Liljencrants, J. (2002). A new approach to intonation analysis and synthesis of Swedish. In *Proceedings of Fonetik 2002*, pages 161–164, Stockholm, Sweden.

Filipsson, M. and Bruce, G. (1997). Lukas - a preliminary report on a new Swedish speech synthesis. In *Working Papers*, number 46, pages 45–56. Department of Linguistics, Lund University.

Fitt, S. and Isard, S. (1999). Synthesis of regional English using a keyword lexicon. In *Proceedings of Eurospeech 99*, volume 2, pages 823–826, Budapest, Hungary.

Fourcin, A. and Abberton, E. (1971). First applications of a new laryngograph. *Medical and Biological Illustration*, (21):172–182.

Frid, J. (1999). An environment for testing prosodic and phonetic transcriptions. In *Proceedings of ICPhS 99*, pages 2319–2322, San Francisco, USA.

Frid, J. (2000). Compound accent patterns in some dialects of southern Swedish. In *Proceedings of Fonetik 2000*, pages 61–64, Skövde.

Frid, J. (2001a). Prediction of intonation patterns of accented words in a corpus of read swedish news through pitch contour stylization. In *Proceedings of Eurospeech 01*, pages 915–918, Aalborg, Denmark.

Frid, J. (2001b). Swedish word stress in optimality theory. In *Working Papers*, number 48, pages 25–40. Department of Linguistics and Phonetics, Lund University.

Frid, J. (2002). Automatic classification of accent and dialect type: results from southern Swedish. In *Proceedings of Fonetik 2002*, pages 89–92, Stockholm, Sweden.

Fujisaki, H. and Hirose, K. (1983). Modelling the dynamic characteristics of voice fundamental frequency with applications to analysis & synthesis of intonation. In *13th International Congress of Linguists*, pages 57–70, Tokyo.

Fujisaki, H. and Hirose, K. (1984). Analysis of voice fundamental frequency contours for declarative sentences of Japanese. *Journal of the Acoustical Society of Japan*, 5:233–242.

Fujisaki, H., Ljungqvist, M., and Murata, H. (1993). Analysis and modelling of word accent and sentence intonation in Swedish. In *Proceeding ICASSP 93*, volume 2, pages 211–214, Berlin, Germany.

Fujisaki, H. and Nagashima, S. (1969). A model for synthesis of pitch contours of connected speech. In *Annual Report Engineering Research Institute*, pages 53–60. University of Tokyo.

Fujisaki, H. and Ohno, S. (1995). Analysis and modeling of fundamental frequency contours of English utterances. In *Proceedings of Eurospeech 95*, volume 2, pages 985–988, Madrid, Spain.

Fujisaki, H. and Sudo, H. (1971). Synthesis by rule of prosodic features of connected Japanese. In *Proceedings of 7th International Congress of Acoustics*, volume 3, pages 133–136.

Féry, C. (1998). German word stress in OT. *Journal of Comparative Germanic Linguistics*, pages 101–142.

Goldsmith, J. (1976). *Autosegmental Phonology*. PhD thesis, MIT.

Grice, M., Baumann, S., and Benzmüller, R. (2003). German intonation in autosegmental-metrical phonology. In Jun, S.-A., editor, *Prosodic Typology and Transcription: A Unified Approach*. Oxford University Press, Oxford.

Grønnum, N. (1992). *The Groundworks of Danish intonation – An introduction*. PhD thesis, Museum Tusculanum Press, University of Copenhagen.

Gussenhoven, C. (2000). Vowel duration, syllable quantity and stress in dutch. Manuscipt, Nijmegen University. Forthcoming (ROA-381).

Gussenhoven, C. (2003). Transcription of Dutch intonation. In Jun, S.-A., editor, *Prosodic Typology and Transcription: A Unified Approach*. Oxford University Press, Oxford. Forthcoming.

Gustafson, J. (1996). *A Swedish Name Pronunciation System*. Speech, Music and Hearing, KTH, Stockholm, Sweden.

Gårding, E. (1977). *The scandinavian word accents*. CWK Gleerup.

Gårding, E. (1983). A generative model of intonation. In Cutler, A. and Ladd, D. R., editors, *Prosody: Models and Measurements*, pages 11–25. Springer, Berlin.

Gårding, E. and Lindblad, P. (1973). Constancy and variation in Swedish word accent patterns. In *Working Papers*, number 7, pages 36–110. Phonetics Laboratory, Lund University.

Haase, M., Kriechbaum, W., Mohler, G., and Stenzel, G. (2001). Deriving document structure from prosodic cues. In *Proceedings of Eurospeech 01*, pages 2157–2160, Aalborg, Denmark.

Hart, J. 't. (1976). Psychoacoustic backgrounds of pitch contour stylization. *IPO - Annual Progress Report*, (11):11–19.

Hart, J. 't. (1991). F0 stylisation in speech: straight lines versus parabolas. *The Journal of the Acoustical Society of America*, 6:3368–3370.

Hart, J. 't., Collier, R., and Cohen, A. (1990). *A perceptual study of intonation*. Cambridge University Press.

Hayes, B. (1987). A revised parametric metrical theory. In McDonough and Plunkett, editors, *Proceedings of the North-Eastern linguistics society*, volume 1, pages 274–289.

Hayes, B. (1995). *Metrical stress theory: principles and case studies*. University of Chicago Press.

Hedelin, P., Jonsson, A., and Lindblad, P. (1987). Svenskt uttalslexikon: 3 ed. Technical report, Chalmers University of Technology.

Heldner, M., Strangert, E., and Deschamps, T. (1999). Focus detection using overall intensity and high frequency emphasis. In *Proceedings of Fonetik 99*, pages 73–76, Gothenburg.

Hieronymus, J., McKelvie, D., and McInnes, F. (1992). Use of acoustic sentence level and lexical stress in HSMM speech recognition. In *Proceedings ICASSP 92*, pages 225–227, San Fransisco, USA.

Hirst, D., Di Cristo, A., and Espesser, R. (1998). Levels of representation and levels of analysis for the description of intonation systems. In Horne, M., editor, *Prosody: Theory and Experiment*, pages ??–?? Kluwer Academic.

Hirst, D. and Espesser, R. (1993). Automatic modelling of fundamental frequency using a quadratic spline function. *Travaux de l'Institut de Phonétique d'Aix-en-Provence*, 15:75–85.

Hirst, D., Nicolas, P., and Espesser, R. (1991). Coding the F0 of a continuous text in french: an experimental approach. In *Proceedings of ICPhS 91*, volume 5, pages 234–237, Aix-en-Provence, France.

Horne, M. and Filipsson, M. (1994). Generating prosodic structure for Swedish text-to-speech. In *Proceedings of ICSLP 94*, volume 2, pages 711–714, Yokohama, Japan.

Horne, M. and Filipsson, M. (1996). Computational extraction of lexico-grammatical information for generation of swedish intonation. In van Santen, J. P. H., Sproat, R. W., Olive, J. P., and Hirschberg, J., editors, *Progress in Speech Synthesis*, pages 443–457. Springer Verlag.

House, D. (1990). *Tonal perception in speech*. Lund University Press, Lund, Sweden.

House, D. and Bruce, G. (1990). Word and focal accents in Swedish from a recognition perspective. In Wiik, K. and Raimo, I., editors, *Nordic Prosody V*, pages 156–173. Turku University.

Jande, P.-A. (2001). Stress patterns in Swedish lexicalised phrases. In *Proceedings of Fonetik 2001*, pages 70–73, Örenäs, Sweden.

Jonsson, A. (1986). A text-to-speech system using area functions and a dictionary. Technical report, Chalmers University of Technology.

Kager, R. (1995). The metrical theory of word stress. In Goldsmith, J., editor, *The handbook of phonological theory*, pages 367–402. Blackwell.

Kiparsky, P. (2003). Fenno-swedish quantity: Contrast in stratal ot. Manuscipt. Forthcoming, http://www.stanford.edu/ kiparsky/Papers/helsingfors.new.pdf.

Klatt, D. (1987). Review of text-to-spech conversion for English. *The Journal of the Acoustical Society of America*, (82):737–793.

Kochanski, G. and Shih, C. (2003). Prosody modeling with soft templates. *Speech Communication*, 39(3–4):311–352.

Kohler, K. (1991). Prosody in speech synthesis: the interplay between TTS and basic research. *Journal of Phonetics*, (19):121–138.

Ladd, D. R. (1996). *Intonational phonology*. Cambridge University Press.

Liberman, M. (1975). *The intonational system of English*. PhD thesis, MIT.

Liberman, M. and Church, K. (1991). Text analysis and word pronunciation in text-to-speech synthesis. In Furui, S. and Sondhi, M., editors, *Advances in Speech Signal Processing.*, pages 791–831. Marcel Dekker, New York.

Liberman, M. and Prince, A. (1977). On stress and linguistic rhythm. *Linguistic Inquiry*, 8:249–336.

Liljestrand, B. (1993). *Så bildas orden*. Studentlitteratur, Lund.

Lindberg, J. (2000). Automatic detection of lexicalised phrases in Swedish. In *Proceedings from Nodalida 99*, pages 103–114, Trondheim. Norway.

Lindberg, J., Lindström, A., and Eineborg, M. (1997). Prediction of word stress in Swedish acronyms – or the difference between a BMX and a BMW. *Proceedings of Fonetik 97. PHONUM*, 4:165–168.

Lindström, A., Horne, M., Svensson, T., Ljungqvist, M., and Filipsson, M. (1995). Generating prosodic structure for restricted and "unrestricted texts". In *Proceedings of ICPhS 95*, volume 2, pages 330–333, Stockholm, Sweden.

Lindström, A. and Ljungqvist, M. (1994). Text processing *within* a speech synthesis system. In *Proceedings of ICSLP 94*, pages 1683–1686, Yokohama, Japan.

Lindström, A., Ljungqvist, M., and Gustafson, K. (1993). A modular architecture supporting multiple hypotheses for conversion of text to phonetic and linguistic entities. In *Proceedings of Eurospeech 93*, pages 1463–1466, Berlin, Germany.

Ljungqvist, M. and Fujisaki, H. (1993). Generating intonation for Swedish text to speech conversion using a quantitative model for the F0 contour. In *Proceedings of Eurospeech 93*, pages 873–876, Berlin, Germany.

Luk, R. and Damper, R. (1996). Stochastic phonographic transduction for English. *Computer, Speech and Language*, 10:133–156.

Lyberg, B. (1981). Some observations on the vowel duration and the fundamental frequency contour in Swedish utterances. *Journal of Phonetics*, 9:261–272.

Malfrère, F. and Dutoit, T. (1997). High quality speech synthesis for phonetic speech segmentation. In *Proceedings of Eurospeech 97*, pages 2631–2634, Rhodes, Greece.

Malmberg, B. (1955). Observations on the Swedish word accent. Report, mimeographed.

Mayo, C., Aylett, M., and Ladd, D. R. (1997). Prosodic transcription of Glasgow English: An evaluation study of GlaToBI. In *Proceedings of an ESCA Workshop: Intonation: Theory, Models and Applications*, pages 231–234.

McCarthy, J. and Prince, A. (1993). Generalized alignment. In Booij and van Marle, editors, *Yearbook of Morphology 1993*, pages 79–153. Kluwer Academic.

Mertens, P., Beaugendre, F., and d'Alessandro C (1996). Comparing approaches to pitch contour stylization for speech synthesis. In van Santen, J. P. H., Sproat, R. W., Olive, J. P., and Hirschberg, J., editors, *Progress in Speech Synthesis*, pages 347–364. Springer Verlag.

Meyer, E. A. (1937). Die Intonation im Schwedischen, I: Die Sveamundarten. *Studies in Scandinavian Philology*, 10.

Meyer, E. A. (1954). Die Intonation im Schwedischen, II: Die norrländischen mundarten. *Studies in Scandinavian Philology*, 11.

Monaghan, A. (1993). The intonation of textual anomalies in text-to-speech. *Xsc*, 12:371–382.

Moulines, E. and Charpentier, F. (1990). Pitch-synchronous waveform processing techniques for text-to-speech synthesis using diphones. *Speech Communication*, 9:453–467.

Möbius, B. (2001). *German and Multilingual Speech Synthesis*, volume 7 of *AIMS*. Universität Stittgart.

Möhler, G. and Conkie, A. (1998). Parametric modeling of intonation using vector quantization. In *Proceedings of 3rd ESCA Workshop on Speech Synthesis*, Jenolan Caves, Australia.

Nespor, M. and Vogel, I. (1986). *Prosodic, Phonology*. Foris, Dordrecht.

Ostendorf, M. and Ross, K. (1997). A multi-level model for recognition of intonation labels. In Sagisaka, Y., Campbell, N., and Higuchi, N., editors, *Computing Prosody*, pages 291–308. Springer-Verlag.

Pagel, V., Lenzo, K., and Black, A. (1998). Letter to sound rules for accented lexicon compression. In *Proceedings of ICSLP 1998*, volume 5, pages 2015–2020, Sydney, Australia.

Pierrehumbert, J. (1980). *The Phonology and Phonetics of English Intonation*. PhD thesis, MIT.

Pierrehumbert, J. B. and Beckman, M. E. (1988). *Japanese tone structure*. MIT Press, Cambridge, MA.

Prince, A. and Smolensky, P. (1993). Optimality theory: Constraint interaction and satisfaction in generative theory. Technical Report 2, Rutgers University Center for Cognitive Science.

Riad, T. (1992). *Structures in Germanic Prosody*. PhD thesis, Stockholm University.

Rischel, J. (1983). On unit accentuation in danish and the distinction between deep and surface phonology. *Folia Linguistica*, 17:51–97.

Ross, K. (1994). *Modeling of intonation for speech synthesis*. PhD thesis, Boston University.

Rossi, M. (1971). Le seuil de glissando ou seuil de perception des variations tonales pour la parole. *Phonetica*, 23:1–33.

Rossi, M. (1978). La perception des glissandos descendants dans les contours prosodiques. *Phonetica*, 35:11–40.

Sakurai, A., Hirose, K., and Minematsu, N. (2002). Data-driven generation of F0 contours using a superpositional model. *Speech Communication*. in press.

Sautermeister, P. and Lyberg, B. (1996). Detection of sentence accents in a speech recognition system. In *The Journal of the Acoustical Society of America*, volume 99, page 2493. (Abstract).

Scheffers, M. (1988). Automatic stylization of f0-contours. In *Proceedings of the 7th FASE Symposium*, pages 981–987, Edinburgh.

Sejnowski, T. J. and Rosenberg, C. (1986). Nettalk: a parallel network that learns to read aloud. *Cognitive Science*, (14):179–211.

Selenius, E. (1972). *Västnyländsk ordaccent*. SLSF 451. Helsinki.

Shih, C., Kochanski, G., Fosler-Lussier, E., Chan, M., and Yuan, J. (2001). Implications of prosody modeling for prosody recognition. In *ISCA Workshop on Prosody in Speech Recognition and Understanding*, Red Bank, NJ.

Shokri, N. (2001). An OT account of Swedish lexical word stress. In Holmer, A., Svantesson, J.-O., and Viberg, A., editors, *Proceedings of the 18th Scandinavian Conference of Linguistics*, volume 1, pages 102–111.

Shriberg, E., Stolcke, A., Hakkani-Tür, D., and Tür, G. (2000). Prosody-based automatic segmentation of speech into sentences and topics. *Xsc*, 32(1–2):127–154.

Silverman, K., Beckman, M., Pitrelli, J., Ostendorf, M., Wightman, C., Price, P., Pierrehumbert, J., and Hirschberg, J. (1992). ToBI: A standard for labelling English prosody. In *Proc. of ICSLP 92*, pages 897–870, Banff, Alberta.

Sproat, R., editor (1998). *Multilingual Text-to-Speech Synthesis: The Bell Labs Approach*. Kluwer Academic, Dordrecht; Boston, MA; London.

Sproat, R. (2000). *A Computational Theory of Writing Systems*. Cambridge University Press.

Sproat, R., Black, A., Chen, S., Kumar, S., Ostendorf, M., and Richards, C. (2001). Normalization of non-standard words. *Computer, Speech and Language*, 15:287–333.

Stolcke, A. and Shriberg, E. (1996). Statistical language modeling for speech disfluencies. In *Proceedings ICASSP 96*, pages 405–408, Atlanta, USA.

Strangert, E. and Aasa, A. (1996). Evaluation of Swedish prosody within the multext-sw project. In *Proceedings of Fonetik 96*, pages 37–40, Nässlingen.

Streefkerk, B. (2002). *Prominence. Acoustic and lexical/syntactic correlates*. PhD thesis, Netherlands Graduate School of Linguistics.

Svärd, N. (2001). Word accents in the Närpes dialect: Is there really only one accent? In *Proceedings of Fonetik 2001*, pages 160–163, Örenäs, Sweden.

Syrdal, A., Möhler, G., Dusterhoff, K., Conkie, A., and Black, A. (1998). Three methods of intonation modeling. In *Proceedings of 3rd ESCA Workshop on Speech Synthesis*, Jenolan Caves, Australia.

Syrdal, A. K., Wightman, C. W., Conkie, A., Stylianou, Y., Beutnagel, M., Schroeter, J., Strom, V., Lee, K., and Makashay, M. (2000). Corpus-based techniques in the AT&T Nextgen synthesis system. In *Proceedings of ICSLP 2000*, Beijing, China.

Taylor, P. (2000). Analysis and synthesis of intonation using the tilt model. *The Journal of the Acoustical Society of America*, 3:1697–1714.

Taylor, P., Black, A., and Caley, R. (1998). The architecture of the festival speech synthesis system. In *Proceedings of 3rd ESCA Workshop on Speech Synthesis*, pages 147–151, Jenolan Caves, Australia.

Taylor, P., Caley, R., Black, A., and King, S. (1999). Edinburgh Speech Tools Library. System Documentation Edition 1.2. Website: http://www.festvox.org/docs/speech_tools-1.2.0/book1.htm, 1994-1999.

Terken, J. (1993). Synthesizing natural-sounding intonation for Dutch: rules and perceptual evaluation. *Computer, Speech and Language*, 7:27–48.

Torstensson, N. (2002). A knowledge-based grapheme-to-phoneme conversion for Swedish. Master's thesis, Department of Computer Science, Högskolan i Skövde.

Touati, P. (1989). SPEAKWORDAC – Swedish Word Accents realized by a text-to-speech system. In *Praktisk Lingvistik 12*, pages 78–81. Institutionen för Lingvistik, Lund Universitet.

Uneson, M. (2002). Burcas – a simple concatenation-based midi-to-singing voice synthesis system. Master's thesis, Department of Linguistics and Phonetics, Lund University.

Van Herwijnen, O. M. and Terken, J. M. B. (2001). Do speakers realize the prosodic structure they say they do? In *Proceedings of Eurospeech 01*, pages 959–962, Aalborg, Denmark.

van Santen, J. and Möbius, B. (2000). A quantitative model of F0 generation and alignment. In Botinis, A., editor, *Intonation - Analysis, Modelling and Technology*, pages 269–288. Kluwer Academic, Dordrecht.

Wang, C. and Seneff, S. (2001). Lexical stress modeling for improved speech recognition of spontaneous telephone speech in the JUPITER domain. In *Proceedings of Eurospeech 01*, pages 2761–2764, Aalborg, Denmark.

Wightman, C. (2002). ToBI Or Not ToBI? In *Proceedings of Speech Prosody 2002*, pages 25–29, Aix-en-Provence, France.

Wightman, C. and Ostendorf, M. (1994). Automatic labeling of prosodic patterns. In *IEEE Transactions on Speech and Audio Processing*, volume 2, pages 469–481.

Willems, N., Collier, R., and Hart, J. 't. (1988). A synthesis scheme for British English intonation. *The Journal of the Acoustical Society of America*, 84:1250–1261.

Wolff, M., Eichner, M., and Hoffmann, R. (2002). Measuring the quality of pronunciation dictionaries. In *Proceedings of PMLA 2002*, Estes Park, Colorado, USA.

Yarowsky, D. (1996). Homograph disambiguation in text-to-speech synthesis. In van Santen, J. P. H., Sproat, R. W., Olive, J. P., and Hirschberg, J., editors, *Progress in Speech Synthesis*, pages 157–172. Springer Verlag.

Zonneveld, W. and Bruce, G. (1999). Word prosody and intonation. In van der Hulst, H., editor, *Word prosodic systems in the languages of Europe*, pages 233–271. Mouton de Gruyter, Berlin and New York.

Öhman, S. (1967). Word and sentence intonation: a quantitative model. *KTH, STL-QPSR*, 2:20–54.

Öhman, S. and Lindqvist, J. (1966). Analysis-by-synthesis of prosodic pitch contours. *KTH, STL-QPSR*, 4:1–6.

TRAVAUX DE L'INSTITUT DE LINGUISTIQUE DE LUND

FONDÉS PAR BERTIL MALMBERG

1. *Carl-Gustaf Söderberg.* A Typological Study on the Phonetic Structure of English Words with an Instrumental-Phonetic Excursus on English Stress. 1959.
2. *Peter S. Green.* Consonant-Vowel Transitions. A Spectrographic Study. 1959.
3. *Kerstin Hadding-Koch.* Acoustico-Phonetic Studies in the Intonation of Southern Swedish. 1961.
4. *Börje Segerbäck.* La réalisation d'une opposition de tonèmes dans des dissyllabes chuchotés. Étude de phonétique expérimentale. 1966.
5. *Velta Ruke-Dravina.* Mehrsprachigkeit im Vorschulalter. 1967.
6. *Eva Gårding.* Internal Juncture in Swedish. 1967.
7. *Folke Strenger.* Les voyelles nasales françaises. 1969.
8. *Edward Carney.* Hiss Transitions and their Perception. 1970.
9. *Faith Ann Johansson.* Immigrant Swedish Phonology. 1973.
10. *Robert Bannert.* Mittelbairische Phonologie auf akustischer und perzeptorischer Grundlage. 1976.
11. *Eva Gårding.* The Scandinavian Word Accents. 1977.
12. *Gösta Bruce.* Swedish Word Accents in Sentence Perspective. 1977.
13. *Eva Gårding, Gösta Bruce, Robert Bannert* (eds.). Nordic Prosody. 1978.
14. *Ewa Söderpalm.* Speech Errors in Normal and Pathological Speech. 1979.
15. *Kerstin Nauclér.* Perspectives on Misspellings. 1980.
16. *Per Lindblad.* Svenskans sje- och tjeljud (Some Swedish sibilants). 1980.
17. *Eva Magnusson.* The Phonology of Language Disordered Children. 1983.
18. *Jan-Olof Svantesson.* Kammu Phonology and Morphology. 1983.
19. *Ulrika Nettelbladt.* Developmental Studies of Dysphonology in Children. 1983.
20. *Gisela Håkansson.* Teacher Talk. How Teachers Modify their Speech when Addressing Learners of Swedish as a Second Language. 1987.
21. *Paul Touati.* Structures prosodiques du suédois et du français. Profils temporels et configurations tonales. 1987.
22. *Antonis Botinis.* Stress and Prosodic Structure in Greek. A Phonological, Acoustic, Physiological and Perceptual Study. 1989.
23. *Karina Vamling.* Complementation in Georgian. 1989.
24. *David House.* Tonal Perception in Speech. 1990.
25. *Emilio Rivano Fischer.* Topology and Dynamics of Interactions - with Special Reference to Spanish and Mapudungu. 1991.
26. *Magnus Olsson.* Hungarian Phonology and Morphology. 1992.
27. *Yasuko Nagano-Madsen.* Mora and Prosodic Coordination. A Phonetic Study of Japanese, Eskimo and Yoruba. 1992.
28. *Barbara Gawronska.* An MT Oriented Model of Aspect and Article Semantics. 1993.
29. *Bengt Sigurd* (ed.). Computerized Grammars for Analysis and Machine Translation. 1994.
30. *Arthur Holmer.* A Parametric Grammar of Seediq. 1996.
31. *Ingmarie Mellenius.* The Acquisition of Nominal Compounding in Swedish. 1997.

32. *Christina Thornell.* The Sango Language and Its Lexicon (Sêndâ-yângâ tî Sängö). 1997.
33. *Duncan Markham.* Phonetic Imitation, Accent, and the Learner. 1997.
34. *Christer Johansson.* A View from Language. Growth of Language in Individuals and Populations. 1997.
35. *Marianne Gullberg.* Gesture as a Communication Strategy in Second Langauge Discourse. A Study of Learners of French and Swedish. 1998.
36. *Mechtild Tronnier.* Nasals and Nasalisation in Speech Production. With Special Emphasis on Methodology and Osaka Japanese. 1998.
37. *Ann Lindvall.* Transitivity in Discourse. A Comparison of Greek, Polish and Swedish. 1998.
38. *Kirsten Haastrup & Åke Viberg* (eds.). Perspectives on Lexical Acquisition in a Second Language. 1998.
39. *Arthur Holmer, Jan-Olof Svantesson & Åke Viberg* (eds). Proceedings of the 18th Scandinavian Conference of Linguistics. 2001.
40. *Caroline Willners.* Antonyms in Context. A corpus-based semantic analysis of Swedish descriptive adjectives. 2001.
41. *Hong Gao.* The Physical Foundation of the Patterning of Physical Action Verbs. A Study of Chinese Verbs. 2001.
42. *Anna Flyman Mattsson.* Teaching, Learning, and Student Output. A study of French in the classroom. 2003.
43. *Petra Hansson.* Prosodic Phrasing in Spontaneous Swedish. 2003.
44. *Elisabeth Zetterholm.* Voice Imitation. A Phonetic Study of Perceptual Illusions and Acoustic Success. 2003.
45. *Johan Frid.* Lexical and Acoustic Modelling of Swedish Prosody. 2003.