

UPPSALA UNIVERSITY
Department of Linguistics

MASTER'S THESIS (20 p)
Language Engineering Programme
Spring Term 2002

Automatic labeling of prosody in Swedish using duration features

Frida Hafvenstein
frida@stp.ling.uu.se

Academic supervisor:
Lars Borin, Department of Linguistics
Uppsala University

Industrial supervisor:
Harald Berthelsen, Babel-Infovox AB

Abstract

There is a constant need for larger prosodically labeled speech databases. Such databases are important to reach a better understanding of prosodic phenomena, and essential for various applications of prosody in the area of speech technology. In this work, efforts to develop procedures for automatic labeling of prosody in Swedish are presented. The main focus is an exploration of the usefulness of duration cues for the signaling of prosodic prominences and phrase breaks in Swedish.

A speech material consisting of about 3 hours and 40 minutes of read speech produced by one female and one male was used as training data. The material was manually labeled for prominences and boundaries. Three levels of prominence, and three levels of prosodic phrase breaks were distinguished in these annotations.

A number of duration features were extracted; including the means of the z-score normalized segment durations across the word, across the stressed and word-final syllables, and the difference between the word-final syllable onsets and rhymes. A decision tree algorithm (CART) was used to classify the prosodic features into the different prominence and prosodic phrase break levels.

The results show that duration cues are indeed important for distinguishing the different levels of prominence as well as of prosodic phrase breaks. The classification of prominence reached a mean recall of 69% and a mean precision of 64% for speaker independent recognition with equal weights given to the three labels. When the hand labeled phrase breaks were used as additional cues to prominence the recall and precision increased to 76% and 78%, respectively. The mean recall and precision for the three phrase break labels reached 50% and 63%, respectively, and when information about prominence label was included the recall and precision increased to 59% and 66%, respectively. However, the results indicate that it would probably be advantageous to try to include additional features, and especially for the automatic labeling of prosodic phrase breaks.

In order to improve the results of the automatic labeling of prominence in Swedish, one would probably need to incorporate other kinds of phonetic information, such as F_0 features, and perhaps also other linguistic information, for example part-of-speech tags.

Acknowledgements

This work has been carried out at Babel-Infovox AB within the frame for the Language Engineering Programme at the department of linguistics, Uppsala University. The people at Babel-Infovox AB have tremendously supported and helped me with all kinds of work, especially my industrial supervisor Harald Berthelsen and the rest of the research group at Babel-Infovox AB: Nikolaj Lindberg, Hanna Lindgren and Jessica Granberg. Mattias Heldner at Centre for Speech Technology (CTT), KTH has also helped me enormously with discussions about phonetic procedures, proofreading and encouragement. Bea Megyesi at CTT, KTH helped me with cross-validation questions. Last but not least I would like to thank my academic supervisor Lars Borin. Thank you all very much!

Contents

CONTENTS	4
1 INTRODUCTION	5
1.1 Purpose	6
1.2 Outline of this thesis	7
2 BACKGROUND	9
2.1 Prosody	9
2.2 Prominence and grouping and their phonetic correlates	11
2.2.1 Prominence	11
2.2.2 Grouping	13
2.3 Prominence and prosodic phrase break labeling	14
2.4 Speech technology systems incorporating prosody	17
2.4.1 Historical background	17
2.4.2 Present systems of automatic prosody recognition.....	18
3 SELECTION OF A MACHINE LEARNING TECHNIQUE	20
3.1 Decision trees and CART	21
4 METHOD	23
4.1 Training and testing with CART using cross-validation	23
4.1.1 Data input to CART	23
4.1.2 Output data from CART.....	26
4.2 Automatic labeling of prominence and prosodic phrase breaks	27
4.3 Evaluation method	28
5 SOLUTIONS AND RESULTS	31
5.1 The prosodic feature vectors to CART	31
5.2 Program development	33
5.2.1 Algorithms and implementations.....	34
5.3 Evaluation and discussion	39
5.3.1 CART results and annotation results	39
6 CONCLUSIONS	50
6.1 Further developments	50
REFERENCES	52
APPENDIX 1: VOCABULARY INDEX	56
APPENDIX 2: PROGRAM CODE	61

1 Introduction

Prosody concerns the rhythmic, dynamic and melodic qualities of speech (e.g. Bruce, 1998), and is used to convey phenomena such as emphasis, intent, attitude and semantic meaning (e.g. Wightman & Ostendorf, 1994). Thus, prosody is of vital importance for spoken human communication. Similarly any speech technology system trying to mimic spoken human communication has to consider the use of prosodic components. Obviously, a speech understanding system lacking prosodic information would not be able to interpret speech the same way humans do, and speech synthesis without prosody would indeed sound monotonous, and hence be more difficult to understand.

A great amount of work in the linguistic literature has been devoted to the theoretical aspects of prosody, and prosodic models have been incorporated in many text-to-speech (TTS) systems. However, current automatic speech recognition (ASR) and automatic speech understanding (ASU) systems typically lack prosodic information (Taylor, 1998; Navas, Hernandez and Ezeia, 2002).

One of the most fundamental functions of prosody is that of marking *prominence* (Bruce, 1998:165). Prominence is used to restrain or highlight units larger than the phoneme, such as syllables, words, or phrases. In doing so one can express differences and similarities in meaning between the units. Another central function of prosody is that of signaling *grouping*, that is the grouping of smaller units (e.g. words) into larger constituents such as prosodic phrases and utterances. Grouping also conveys the boundaries between these elements (Bruce, 1998:124).

The company Babel-Infovox AB, where this master's thesis was carried out, needs labeled speech with information about prominence and grouping for their unit selection speech synthesis. This is because the unit selection method picks out the best fitting units, e.g. phones, syllables and words, including the prominence level and grouping information on that unit. At this moment, the information about prominence and grouping is hand labeled at Babel-Infovox AB. As this manual labeling is very

time-consuming and therefore also expensive, automatic labeling of prosody would make the work considerably more economical (cf. Campbell, 1993:344).

Furthermore, automatic labeling of prominence and grouping can provide large amounts of labeled speech material, which can be analyzed with a higher statistical reliability. This can contribute to a better understanding of the nature of prosody and help forming better prosodic models for speech synthesis. Automatic labeling can also be used in speech understanding systems in order to improve the interpretation of the speech (Wightman & Ostendorf, 1994).

This report forms a part of a master's thesis in language engineering with a specialization in speech technology ranging one term. The work was carried out at the speech technology company Babel-Infovox AB in Stockholm, within the framework of the Language Engineering Programme at the Department of Linguistics at Uppsala University.

1.1 Purpose

The purpose of this thesis is to develop procedures for automatic labeling of prominence and prosodic boundaries in Swedish. These procedures will be based on quantitative data analysis and classification procedures using machine learning techniques. Although it is a known fact that several acoustic correlates, including segment duration, fundamental frequency (F_0) and energy, interact to convey prominence and prosodic boundaries, this thesis will focus on the usefulness of segment duration as input to a prosodic classifier. Duration cues are chosen as a starting point partly because they are known to be important for the signaling of prominence and boundaries, and partly because this information is already accessible in the segmentation of speech databases without any further access to the speech waveform or any special signal processing (Campbell, 1993:344). The area of automatic labeling of prosody is relatively unexplored. Automatic labeling using a

rule-based system for prosodic modeling has been explored for Swedish (e.g. House & Bruce, 1990). Machine learning techniques for prosodic classification have however, to my knowledge, not been used for Swedish, and only to a small extent in other languages.

The segmental information in this case is available in speech databases collected at Babel-Infovox AB. In addition, these databases contain the manually labeled prominences and prosodic phrase boundaries, which will serve as a predictive model to the classifier. This speech database provides the only input data to the classification model. The output will be a similar speech database but with automatically labeled prominence and prosodic phrase break levels for each word.

For the classification of prosody an appropriate machine learning technique will be chosen. The classifier will produce probabilistic information of the spread of the feature values into the given classes (prominence and boundary levels). The feature vector, which is input to the classifier, will be selected based on earlier works on prosody labeling and to some extent on experimental results. A method to automatically extract features will also be explored and implemented. The probabilities from the classifier will then be implemented in the prosody labeler which transcribes the most probable prominence level and prosodic boundary level into the database. The results of this annotation compared to the results from the classifier will be evaluated both qualitatively and quantitatively for all test data.

1.2 Outline of this thesis

The development of procedures for automatic labeling of prosody requires knowledge from several areas. Therefore, the Background chapter (2) contains brief accounts covering a number of areas, i.e. prosody, classification and labeling of prosody, state-of-the-art of prosody and machine learning. These areas are outlined in the following way: First a short introduction of prosody is given in 2.1, to continue with a more in-

depth account of the prosodic functions prominence and grouping (2.2). A section on issues in the classification and labeling of prosody follows in 2.3, and section 2.4 ends chapter two with a short outline of the background and state-of-the-art of prosody incorporated in speech technology systems.

In chapter three the selection of a machine learning technique is reported. First follows a short description of the characteristics and main advantages of the machine learning technique used for the classification of prosody in this thesis: decision trees (3) compared to other important machine learning techniques that have been used in similar tasks. This chapter should have been a section in the chapter about solutions and results since choosing a machine learning technique is a part solution to the task of developing procedures for automatic labeling of prosody. The reason for this structure is simply that the information given in this chapter is needed when the Method chapter (4) is read.

The Method chapter (4) is a report on the methods for collecting data, training and testing with the decision tree algorithm CART and how the output from CART is used to automatically label the speech data. The evaluation method is also accounted for in Method. Before a report of the evaluation results a description is given of the solutions to the implementation problems of the computer programs for the automatic labeling (5). In Conclusions a summary of the most important observations made in this work and suggestions for further developments of this work are given (6).

The target group of this thesis is the students who have read all compulsory courses at the Language Engineering Programme at the Department of Linguistics at Uppsala University. Because the course Phonetics and phonology is a compulsory course in the program, terms from phonetics and phonology, and some terms from the speech technology area will be used without any definitions or explanations in the text. However, because some terms can be forgotten easily and also in order to be able to reach more people than just the target group, a vocabulary for the most important terms in this thesis, which also functions as a small index to the thesis, is appended (Appendix 1).

2 Background

2.1 Prosody

The study of prosody has gained increasing interest from phoneticians and phonologists during the last decades. One reason for this is probably the possibilities for analyzing larger speech materials given by the technical progress of computers in the instrumental analysis of speech (Bruce, 1998:10).

A change in point of view in the main theories of phonology was also an important contribution to the interest of prosody (Bruce, 1998:10). The new phonological theories are collectively referred to as *non-linear phonology* of which *autosegmental phonology* and *metrical phonology* are the two major theories in this field. These theories abandon the segmental view represented by *Sound Pattern of English* by Noam Chomsky and Morris Halle, with a strict sequential analysis of speech sounds, i.e. vowels and consonants (Gussenhoven & Jacobs, 1998). Instead, the emphasis is on prosodic phenomena, such as stress and rhythm and tone and accents (Gussenhoven & Jacobs, 1998). A third motivation for the great interest of prosody is the need for better speech technology applications (Bruce, 1998:10).

The increased interest in prosody calls for descriptions of the phenomena and of the relevant prosodic categories, descriptions that has proven difficult to give (Bruce, 1998:10). As mentioned above, the non-linear phonology has a non-segmental perspective; a *suprasegmental* perspective, which is a description of the speech above the segmental level, i.e. above the vowels and consonants. Prosody is often described as being the phenomenon in speech that are represented on the suprasegmental level, i.e. that the domain of prosody is larger than the segment.

The prosodic categories can be described by looking at what they contribute with in speech. A common view is that prosody gives structure to speech in order to convey emphasis, semantic meaning, intent and attitude (e.g. Wightman & Ostendorf, 1994; Bruce, 1998). These expressions belong to the functions of prominence,

grouping and different discourse functions, which are the most important functions of prosody according to Bruce (1998:15). But prosody also has a distinctive function, e.g. *quantity* in Swedish. In Swedish, quantity reaches from the stressed vowel to the following consonant or consonant cluster, i.e. the rhyme of the syllable, e.g. /tak:/ (tack = thank you) vs. /ta:k/ (tak = roof/ceiling) (see Figure 1).

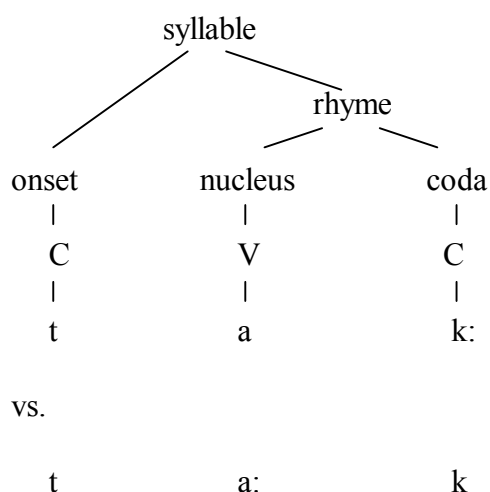


Figure 1. An example of a syllable structure and the difference in quantity between the Swedish words *tack*, /tak:/ (thank you) and *tak*, /ta:k/ (roof/ceiling).

Prosody is often associated with the rhythmical, dynamical and melodic qualities of speech, and hence linked to the acoustic phonetic dimensions: duration, intensity and pitch (F_0) (e.g. Bruce, 1998:11). However, this is not the complete picture of how prosody is conveyed. A number of other phonetic correlates, e.g. voice quality and breathing, are probably also involved in the signaling of prosody. In fact, Bruce states that all acoustic phonetic dimensions are involved in the signaling of prosody (Bruce, 1998:11).

The fact that what we perceive as prosody has certain acoustic correlates makes it possible to investigate the mappings between those dimensions. Statistical methods have proven successful in the processing and analyzing of large amounts of prosodic labeled speech data (e.g. Campbell, 1993; Wightman & Ostendorf, 1994; Shriberg et

al., 2000; Greenberg, 2002). Greenberg explains harshly but with, according to my opinion, some veracity, the importance of the study of real speech data:

Sophisticated technology depends on “getting the details right” to a degree that far exceeds what passes for knowledge and insight within the domain of science [...]. With respect to speech the contrast between “pure” and “applied” research is stark indeed. Linguists and phoneticians often view spoken language through a “glass menagerie” of abstract forms, which often bear only the faintest resemblance to language spoken in the “real” world. (Greenberg, 2002:4)

I would say that linguists and phoneticians are beginning to use more corpus-based approaches but that the view Greenberg speaks of still exists.

2.2 Prominence and grouping and their phonetic correlates

2.2.1 Prominence

Prominence is expressed with phonetic correlates such as segmental lengthening, higher intensity, spectral contrasts and bigger F_0 excursions (Bruce, 1998:42, Wightman & Ostendorf 1994:470). The size of the F_0 movements discriminates between degrees of accented words and accented vs. unaccented (stressed) words, but not between stressed and unstressed syllables (Bruce, 1998:42). The vowel quality is also affected in the signaling of prominence. When a word is in focus, i.e. has high prominence, the phonemes in the word are lengthened and when there is more time for the articulation, the gesture becomes more extreme, with the formants in the sound spectrum more spread apart (Bruce, 1998:71). This means that one could also study the phoneme qualities to detect or improve detection of prominence which Greenberg actually has done (see section 2.4.2 for more information about his work) (2002).

Levels of prominence are relatively expressed with degrees of phonetic correlates (Elert, 1991:118) in the way that a stressed syllable for example is expressed with different degrees of stress depending on the context, speaker variations and the

position in the phrase (Strangert & Heldner, 1998). The relative importance of these phonetic correlates and the interaction between them is however not yet completely understood (Wightman & Ostendorf, 1994:469).

There is no general agreement across languages as to the number and definitions of prominence levels. But according to Bruce, who has made important contributions in describing the prosody of Swedish, there are, not counting the category of unstressed, three linguistically relevant prominence levels in Swedish: stressed, accented and focused. These categories were agreed upon for basic labeling of prominence in Swedish. It is part of a minimal phonological prosodic transcription system developed within a language engineering research program carried out in the years of 1990-1996 which is based on IPA (International Phonetic Alphabet) and includes categories for grouping as well (see chapter 2.2.2).

It is commonly known that prominence is partly expressed by lengthening. A study of Strangert and Heldner, for example indicates a lengthening of focused words of 25% in general compared to non-focused words (1998:3305). Fant with colleagues conclude that “[...] duration is the most robust correlate, whereas the importance of F_0 and intensity measures vary with respect to position and prosodic grouping.” (Fant et al., 2001:4). The domains of prominence, i.e. the unit affected by the acoustic phonetic signals, are the syllable, foot and word for stressed, accented and focused respectively, according to Bruce (1998:80). Heldner and Strangert refine the domain of the focal accent to the stressed syllable plus the following syllable (their study concerns non-compound words) (Heldner & Strangert, 2001).

Further, the segments in the stressed syllable are lengthened differently so that the contrast between the short and long segments in a stressed syllable is enhanced (Heldner, 2001:103). The consonants in the onset and coda are lengthened irrespective of their inherent length whereas short vowels may not be lengthened when the word is focused (Heldner & Strangert, 2001:351). This is to retain a complementary relation between the vowel and the consonant in the same

morphological syllable so that V:C and VC:¹ have nearly the same length (Fant, Kruckenberg and Nord, 1990:71).

2.2.2 Grouping

Prosodic boundaries and groupings are very important for spoken communication. By grouping words into prosodic constituents of varying sizes, and indicating boundaries between these constituents, the speaker facilitates the listener's processing of the message. This is done by means of the prosodic function grouping (Bruce, 1998:124). The different prosodic groups are sometimes linked to the morphological or syntactic equivalent, especially when we are speaking of syllables and words. But when it comes to compounds or bigger groups than words, like phrases and utterances the difference between syntax and prosody is greater and also the speaker variation increases (Campbell, 1993:351).

Since the prosodic constituents are of different sizes, there are also different degrees of boundaries between the groups. But the categorization of these degrees of boundaries and how the different categories are characterized diverges between different research projects and languages. For Swedish, the basic prosodic labeling, mentioned in previous section, (2.2.1) has three prosodic boundary categories for grouping, excluding the category of no boundary: prosodic phrase, prosodic utterance and speech paragraph (Bruce, 1998:166).² Concerning the phenomenon of grouping, the task for this essay only includes boundary labeling of the prosodic utterance and the prosodic phrase to use the Swedish basic prosodic transcription labels.

Prosodic grouping is signaled through several acoustic phonetic correlates such as F_0 , duration, intensity, voice quality and pauses (Bruce et al., 1992:113). Studies of how the F_0 curve behaves in prosodic grouping are thorough while there is not

¹ V:C = long vowel (V:) and short consonant (C) and VC: = short vowel (V) and long consonant (C:).

² The two former categories are, for example comparable with number three (3) and four (4) respectively in the 5-level break index ToBI (Tone and Break Indices) (Beckman & Ayers Elam, 1997) and Pierrehumbert's Intermediate phrase and Intonational phrase (Perrehumbert, 1980).

enough knowledge of how duration, pauses, (Horne, Strangert and Heldner, 1995:170) intensity and voice quality work in prosodic grouping (Bruce et al., 1992:113).

However, Horne with colleagues have worked on how duration works in grouping and they support the assumption of the phenomenon of *final lengthening* in Swedish (Horne et al., 1995). Final lengthening is when the last part of a speech unit is lengthened due to grouping. In this small study the boundary categories were the same as those in the basic prosodic labeling for Swedish and the only word preceding the boundaries was the Swedish word “procent” (percent) tested with both focal and non-focal accent. They found that the final lengthening occurred before the prosodic phrase and the prosodic utterance. The study indicates that the domain of the lengthening in Swedish is the last segment in the coda of the last syllable (Horne, Strangert and Heldner, 1995:173). This could be different for different languages. Campbell says for example that final lengthening affects the entire coda of the final syllable (Campbell, 1993:348). Final lengthening occurs both in focally accented words and in non-focal words in Swedish (Horne, Strangert and Heldner, 1995).

Likewise the initial syllables of an utterance can also be lengthened by grouping, which means that the position in the phrase affects the syllable duration (Heldner & Strangert, 2001:345). Moreover, silent intervals occur mostly at the prosodic phrases and the prosodic utterances (Horne, Strangert and Heldner, 1995:173, Bruce 1998:150).

2.3 Prominence and prosodic phrase break labeling

A prerequisite of both manually and automatic labeling of prosody, is solutions to three tasks: a linguistic interpretation of the relevant changes in the physical properties of the speech wave, a classification of these changes into discrete categories, and finally a design of the notational system representing the prosodic categories. These

tasks are complex and difficult matters. As described above, it is not fully understood how the physical properties of the speech wave that corresponds to a speaker's utterance behave or how they interact in order to convey prosodic information (see section 2.2.1). Furthermore, speakers differ in their way of expressing prosody and the influence of the speaking style can also be interpreted as the prosody, e.g. if the speaker utters every word with strong emphasis, every word can erroneously be labeled as having high prominence. In manual labeling the linguistic interpretation is usually done perceptually, however, and is not dependent on knowing the acoustic phonetic signal.

The classification is also a complex matter. One important question concerning this task is how to treat the continuous speech signal: as such or as discrete. If it is classified into discrete values, then one has to decide how many levels there should be within one category, e.g. prominence. These choices depend on several matters, e.g.:

- The properties of the speech, e.g. if it is multilingual or not.
- The function of the labeled speech, i.e. if it is used in an automatic speech understanding (ASU) system, in a text-to-speech synthesis system (TTS) or for some other purpose.
- The procedure of the classification, i.e. if it is a theory-driven system or a data-driven system.
- The economy

The design of the notational system also differs depending on the context of labeling, i.e. if the labeling is done automatically or by hand, if it is to be done computer-readable or not and if it is a multilingual vs. non-multilingual system.

There are a few attempts at a standard labeling notation, e.g. ToBI (Tone and Break Indices), INTSINT (International Transcription System for Intonation), TILT, IPA (International Phonetic Alphabet), and SAMPROSA (SAM Prosodic Alphabet) (Wells, 2003) even though they do not have exactly the same purpose or fill the same needs (Llisterri, 2003). ToBI was developed for use in speech research and speech technology for American English, but has been extended to other languages as well.

ToBI consists of four parallel tiers: an orthographic tier, a break-index³ tier, a tone tier and a miscellaneous tier. Regarding the tone tier, the problem with systems like ToBI is, according to Batliner et al., that they “[...] only introduce a quantization error” in the labeling of prominence (2001b:2). They reduce the acoustic values of F_0 into almost binary labels but add no phonological or linguistic value to it. The labels for F_0 /pitch only give information about the acoustic phonetic signal and not about which prosodic phenomenon it conveys. These systems are also too focused on the contribution of F_0 in comparison to other prosodic features (Batliner et al., 2001b:2).

INTSINT was designed to provide a cross-linguistic comparison of prosodic systems for the symbolization of intonation. ToBI and INTSINT could be good systems for a phonological description but not easily applicable in speech technology systems according to Batliner et al. (2001b:2).

TILT is yet another system for transcription of intonation events, like pitch accents or boundary tones⁴ (Taylor, 1998). Labels for prosodic phrase breaks are however phonologically more justifiable in the transcription systems mentioned above.

IPA and SAMPROSA are also intended both for prosodic transcription for linguistic analysis and for speech technology. These two systems have prosodic labels for both an intonational tier and a more abstract representation of prominence and grouping.

Other important works on the mapping of prosody onto phonetic correlates are those of Fant with colleagues (Fant, Kruckenberg and Liljencrants 2000) and Portele & Heuft (1997). Fant with colleagues have mapped phonetic correlates to prominence levels in Swedish with the help of evaluators marking on a scale from 0 to 30 how prominent they perceive a syllable. In that way one can see which phonetic correlates are the most important in the perception of prominence and what phonetic values the phonetic prominent syllable have (Fant, Kruckenberg and Liljencrants, 2000). They

³ A break-index is an index describing the different prosodic boundaries.

⁴ Boundary tones are the tones in a sentence conveying whether it is a statement or a question (Wightman & Ostendorf, 1994:470).

studied not only the duration and F_0 correlates but also other prosodic parameters that concern the whole speech production e.g. the breathing, the voice source and the segmental composition. Portele & Heuft have used a similar system as Fant & Kruckenberg but complemented it with speech synthesis rules (1997).

When solutions to the issues mentioned above are provided, the actual labeling can be carried out. According to Krippendorff (1980) there are three issues concerned in the evaluation of a labeling system: stability, reproducibility and accuracy. Stability is how consistent the labeling is between the same transcriber at two different times. Reproducibility is how two or more transcribers agree about the classification labels, and accuracy is how many labels that are correct compared to a template with correct labels. Results from hand labeling, however, indicate that these criteria are difficult to comply with (Strangert & Heldner, 1994). In automatic labeling high stability and reproducibility are obviously easier to reach assuming that the labeling is done through executing the same computer program at different times and that the program then does the same thing if it gets the same input.

2.4 Speech technology systems incorporating prosody

2.4.1 Historical background

Automatic labeling of prosody is about automatic recognition of prosody. In the beginning of the 1980's the automatic recognition of prosody did not work very well because neither the recognition technique nor the phonological theories of prosody were well developed (Wightman & Ostendorf, 1994:469). The earliest approaches to recognition of prosody have been based on F_0 analysis using linear regression, HMM-based algorithms and dynamic programming to find the best prosodic phrase segmentation (Wightman & Ostendorf, 1994:470). Wightman and Ostendorf have concluded that duration cues are good complements to F_0 information in the analysis

of at least grouping. This fact was not known when the first systems were created (Wightman & Ostendorf, 1994:470).

Energy also appears to be an important cue for automatic detection of prominence. Little discussion of energy cues can be found in the linguistic literature, however, probably because energy is less important than F_0 and duration in human perception of prominence. However, focal accentuation has been analyzed later on using energy and duration features in HMM's and linear discriminant functions⁵ (Wightman & Ostendorf, 1994:470).

The speech understanding systems so far are in general intended for restricted tasks with short utterances where a prosodic segmentation would not help much which is why prosodic recognition to such a small extent has been incorporated in the systems (Nöth et al., 2000:519).

2.4.2 Present systems of automatic prosody recognition

There are few systems handling automatic recognition of prosody today and to my knowledge none of them are for Swedish. The majority of them use prosodic feature vectors with values of acoustic phonetic features like segment duration, syllable duration, pause duration, energy, voice quality, different F_0 information like F_0 contours, F_0 maxima and minima, etc. Wightman & Ostendorf for example use prosodic feature vectors within decision trees and a Markov model in recognizing prosodic grouping and phrasal prominence in English at a post-word-recognition (1994). Another project that uses the feature vector method is Verbmobil, which is a translation system with English and German among the translated languages (Batliner et al., 2001a; Nöth et al., 2000). A Linear Discriminant Analysis, (LDA) was used as predictor of a large amount of prosodic features (Batliner et al., 2001a:23). The most important features for both English and German and for both prosodic boundary and

⁵ A linear discriminant function or Linear Discriminant Analysis (LDA) is a type of classification technique.

accent classification were those modeling duration, in combination with energy, followed by pauses and F_0 (Batliner et al., 2001a:23). As Batliner et al. conclude:

(...) [I]t seems that the 'prototypical' prosodic feature (group) F_0 is not that important for the prediction of these two 'classic' prosodic events, i.e. the marking of boundaries and accents. (Batliner et al., 2001a:23)

In another study within the Verbmobil project a Multilayer Perceptron model (MLP, a kind of neural network) is found to work best for their data in classifying prominence in comparison to Gaussian distribution classifiers (Nöth et al., 2000:524). Andrew Hunt has developed a similar system using both prosodic and syntactic features in the vector analysis (Nöth et al., 2000:520). Another work that uses features is that of Campbell. He has automatically detected prosodic phrase boundaries and prominence using only duration features (1993).

TILT is yet another example of a system based on feature vectors. It automatically analyzes intonation using Neural Networks and HMM's (Taylor, 1998:9). Intonation is here represented as sequences of events, like pitch accents or boundary tones. The events have the acoustic parameters amplitude, duration and *tilt* (a measure of the shape of the F_0 -curve) (Taylor, 1998:5). Taylor distinguishes between prosodic detectors and prosodic classifiers where TILT is a prosodic detector of the acoustic phonetic values correlated to prosodic events while Wightman & Ostendorf and Verbmobil have developed prosodic classifiers which classify the prosodic events.

Another work on prosodic detection is that of Greenberg, called AutoSAL (2002). The AutoSAL system has, as successfully as human transcribers, labeled "stress-accent patterns" using MLP (2002). Greenberg has also concluded, like Batliner et al. that using F_0 features are not that contributive to stress accents as it has been claimed. He found that, out of all 45 combinations of phonetic features, the most important ones for prosodic detection were vocalic identity, duration and energy (2002).

3 Selection of a machine learning technique

As explained above, prosody is a very complex and difficult phenomenon, which is not fully understood. This, obviously, makes learning about prosody and its phonetic correlates very difficult. But, with the high processing and memory capabilities of a computer, the learning algorithm can be much more effective and process large amounts of data. Instead of drawing conclusions about which phonetic correlates that interact in order to convey prosody and what values they have from small sets of speech data that are arranged to suit the analysis, a machine learning program can present statistically reliable information about actual instances.

The learning task in this thesis is a classification problem (Mitchell, 1997:54) for prominence and grouping. The task is to find a learning algorithm that systematically predicts discrete classes for new unseen data based on *instances*, i.e. words, with already hand labeled classes; *concepts* and relevant duration data; *attributes*. Such a learning procedure is called supervised, i.e. the learning has as input a scheme with actual instances (Frank, 2003).

The most common machine learning techniques in classification of prominence and prosodic phrase breaks seems to be the decision tree algorithm CART and MLP (Multilayer Perceptrons) (e.g. Wightman & Ostendorf, 1994, Shriberg et al., 2000; Nöth et al., 2000; Greenberg 2002). Other machine learning algorithms used for this purpose are LDA and Gaussian distribution classifiers (e.g. Batliner et al., 2001a). CART and MLP often produce classification results of comparable accuracy (Mitchell, 1997:85) whereas the latter techniques do not give as good results as CART and MLP (Nöth et al., 2000; Batliner et al., 2001a). Therefore only a comparison between CART and MLP has been made in the selection of the machine learning technique for this thesis.

Compared to MLP, the CART algorithm is faster in the training time. The training time of a MLP can range from seconds to hours, depending on several factors, e.g. the number of training examples considered and the settings of the different learning

parameters for the MLP (Mitchell, 1997:85). Moreover, CART gives more understandable and interpretable output information about the class predictions made from the data than a MLP does (Mitchell, 1997:54). These two advantages and the fact that MLP and CART often give as good results, are the main reasons for choosing the learning algorithm CART for this work. There are of course limitations of how much time the training of the data can take in this work, so it is important that the training method is fast. The understandability and interpretability of the output of CART are also important in the modeling of prosody. The properties and some more advantages of CART are described below.

3.1 Decision trees and CART

The most practical and commonly used machine learning technique for learning based on experience, inductive inference, is according to Mitchell decision trees (1997:52). The method has been used successfully in the work of classifying prosody (Wightman & Ostendorf, 1994, Shriberg et al., 2000).

Decision trees are probabilistic classifiers that sort the instances in the learning data by binary questions about the attributes that the instances have. It starts at the root node and continues to ask questions about the attribute of the instance down the tree until a leaf node is reached (Mitchell, 1997:52). The decision tree algorithm selects the best attribute and question to be asked about that attribute at each node. The selection is based on what attribute and question about it divide the learning data so that it gives the best predictive value for the classification. When the tree has reached the leaf nodes the probability about the class distribution of all instances in the corresponding branch is calculated which is used as predictors for new unseen test data. (Shriberg et al., 2000)

The selection of the node splitting questions is based on an information-theoretic criterion called entropy, which is a measure of how much information some data

contains. In the decision tree, entropy can be measured by looking at how “pure” the resulting subsets of a split are, i.e. if a subset contains only one class it is purest, and the opposite, the largest impurity is hence when all classes are equally mixed in the subset (Breiman et al., 1984).

Furthermore the decision tree algorithm has a criterion for when to stop splitting the nodes to avoid overfitting the tree to the training data. Overfitting is when the tree makes decisions based on too small subsets with data that does not represent the whole data structure. This leads to erroneously classified test data. (Breiman et al., 1984; Manning & Schütze, 1999:582). The stopping criterion is often based on entropy as well. The criterion could for example be to stop splitting the nodes when the purity of the node is the highest (Manning & Schütze, 1999:583). The most common and according to Breiman et al. the best procedure, though, is called *pruning*, which means that the tree first is grown large and then cut back, pruned, to the best size (Breiman et al., 1984; Manning & Schütze, 1999).

There are several decision tree algorithms with diverging properties and advantages. The decision tree algorithm CART is probably the most important contribution to the field of theoretically founded decision trees. It basically extends the decision tree method, to handle numerical values and is able to incorporate regression trees when needed. CART is robust to noisy data, can handle missing data and non-homogeneous attributes and do not require assumptions about the independence of the concepts (Mitchell, 1997:52; Wightman & Ostendorf, 1994:473). In addition, this output can be compactly stored and classify new data efficiently (Breiman et al., 1984).

4 Method

4.1 Training and testing with CART using cross-validation

4.1.1 Data input to CART

The only information source used for training and classification with CART is speech databases written in XML (Extensible Markup Language). There were 9 speech databases available at Babel-Infovox suitable for this task. The speech is in total about 3 hours and 40 minutes of read news sentences where five of the nine databases are read by a female and four of them are read by a male. In Table 1 the numbers of syllables and words in each database and the total number of the syllables and words of all nine databases can be seen.

Table 1. Number of syllables and words for each speech database and the total number in the last row.

Speech databases:	No of syllables:	No of words:
1, female	3806	2220
2, — " —	4619	2568
3, — " —	5519	2932
4, — " —	6757	3531
5, — " —, (weather report)	4114	1987
6, male	2421	1431
7, — " —	4628	2575
8, — " —	5518	2932
9, — " —, (weather report)	4148	1987
Total number	37416	20176

The speech databases are automatically segmented into segments, (including phones and pauses) syllables and words. The segment duration and the stress degree of the syllable are also automatically detected. After the segmentation and transcription of the audio signal the XML databases are manually labeled with respect to prominence

and prosodic breaks and checked by another human transcriber. An excerpt of a speech database is shown in Figure 2 below.

```
<token orthography="Det" type="word" id="2">
  <reading prominence="deacc"
    part_of_speech="PN_NEU_SIN_DEF_SUB/OBJ"phrasing="0"
    lemma="det" word_transcription="d'Et" comment="">
    <transcription type="manual">
      <syllable accent="1">
        <phoneme symbol="d" duration="50"/>
        <phoneme symbol="E" duration="135"/>
      </syllable>
    </transcription>
  </reading>
</token>
```

Figure 2. Excerpt from one of the speech databases.

Different duration information, pause information, information about syllable stress and the prominence and prosodic phrase break levels for each word are used as input to the classifier. The CART software used in this task, Wagon, from the Centre for Speech Technology Research, University of Edinburgh, requires a special structure of the input, a *feature vector* (Taylor et al., 2003). Each row of the feature vector represents one sample containing one *predictee* and one or more *predictors*, each row having the same order of the predictors with the predictee as the first element in the row (e.g. Figure 3).

```
foc 0.328 2.341 1.267 pause 1 0 0 1
acc 2.439 4.562 nil nopause 0 1 0 1
acc 1.744 2.575 2.266 pause 0 0 1 1
(etc.)
```

Figure 3. Example of a feature vector with the predictee first on each row and the predictors in the same order on every row.

In the prosodic feature vector (PFV) in this work, one row represents one word and the predictee is the prominence level or phrase break level for the word in question. The predictors are values of different prosodic cues that are chosen. Thus, the

information from the speech databases will be extracted and prepared with a computer program designed and implemented in this work (see section 5.2). In addition a description file is needed to describe the feature values in the PFV, i.e. what kind of information the features should accept, e.g. if it is real numbers, binary values etc.

The three prominence levels in the speech database are, deaccented (*deacc*), accented, (*acc*) and prominenced (*prom*), though deaccentuation is not a marking of prominence but a marking of the lack of the higher prominence level, accented. The division of prominence corresponds to Bruce's prosody model with the three levels stressed, accented and focused (Bruce, 1998:80). The prosodic phrase break levels are no break, (*NB*) small break (*SB*) and big break (*BB*). These phrase breaks are comparable to the prosodic phrase boundary and the prosodic utterance boundary respectively that are agreed upon for prosodic labeling in Swedish (see section 2.2.2) (Bruce, 1998:166).

The selection of the predictors is mainly based on discussions with a phonetician on what duration features are possible and reasonable to use, taking into consideration duration features used in other works on automatic labeling of prosody. These predictors and motivations for them will be presented in section 5.1 in the next chapter, Solutions and results (5). The duration of the features is normalized with respect to the different intrinsic duration of the phonemes, the differences in speaking rate and the way prosody is expressed. The duration is normalized so that a comparison of how much longer or shorter the duration is than expected can be done (Wightman et al., 1992:1711; Campbell, 1993:345). The normalized duration of a segment, i.e. the z-score of the segment's duration, is a measure of the number of standard deviations from the mean duration of the phoneme that segment is a token of, i.e.:

$$z\text{-dur} = [d(i) - \mu_p] / \sigma_p$$

The $d(i)$ is the duration of the segment, the μ_p is the mean value of all segments' duration of the phoneme that $d(i)$ is a token of and σ_p is the standard deviation of the phoneme duration.

The training of CART will be done on both speaker-independent and -dependent speech data, to test which method gives the best results.

A training- and testing-technique called *k-fold cross-validation* is used due to the small amount of available speech data for training and testing of the classification. All speech databases can then be used both as training and test data without testing on the same data as CART is trained upon at that moment. Firstly, the training material is divided into k subsets, then each subset, in turn, is used for testing and the remainder for training. Sentences are randomly extracted from the speech databases to form k subsets with similar distribution regarding the values of the predictors and predictees.

4.1.2 Output data from CART

Because the training and testing will be done through cross-validation there are k output files, trees, with classification results for prominence and phrase breaks. The tree size will be optimized by means of a stopping criterion. The stopping criterion is a decision of the numbers of words that is the minimum for a question, a node, in the tree. This number is reached by testing which stopping number gives the best result with respect to the classification rate given by Wagon.

The classification rate given by Wagon does not reveal whether it is a measure of the number of correctly classified labels compared to the number of all correct labels, i.e. recall, or a measure of the number of correctly classified labels compared to the number of all classified labels, i.e. precision. However, because the number of all classified labels, in this case, is the same as the number of all correct labels, the two measures would give the same rate. Hence, information about what measure Wagon outputs, recall or precision, would not change the procedure of finding the best stopping number for the tree.

The search pattern for finding this stopping number is a tree starting with the default number, 50 and increasing or decreasing with 10 and at the leaf nodes with 5 (see Figure 4).

The reason for not choosing the better and more common stopping method, i.e. pruning, is simply that it is a more time consuming method, which this work does not allow due to limited resources.

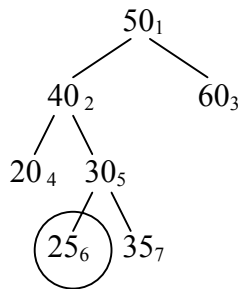


Figure 4. An example of a search for the best stopping number. The lowered figures show the order of tests, i.e. the first test is a tree with 50 as the stopping number and the second is with 40 and so forth. The ring is the best stopping number in this search.

The classification results will then be evaluated and compared to the results from the actual labeling of the XML databases. The best tree out of the k trees is chosen for the labeling process by means of averaging the stopping values for all the trees. Then a new tree is grown on all subsets with the averaged stopping value (Manning & Schütze, 1999:586). The predicted classification is then used in the labeling process, but because the output from Wagon does not show the predicted labels a tree parser is executed to extract the information.

4.2 Automatic labeling of prominence and prosodic phrase breaks

The labeling of prominence and phrase breaks are done separately so it is possible in the future to make a choice of what prosodic information shall be labeled. The actual labeling is done with a simple program that takes the predicted labels from CART and writes them onto manually labeled XML databases so that the evaluation can be done

by a comparison between the manually labeled and the automatically labeled prominence and phrase breaks.

4.3 Evaluation method

Evaluation will be done both on the outputs from CART, i.e. the classification results, and on the labeling results. In the classification of three categories a 2×2 contingency table is made for each category, C_i separately (evaluating C_i versus $\neg C_i$), viz. in Table 2. The cases where e.g. deacc is correct and also assigned deacc are called *true positives* in the contingency table. *True negatives* are the cases where any of the other categories is the correct label and also assigned any of the other categories. The other two figures in the contingency tables are *false positives* and *false negatives* where the former is all cases that the system erroneously assigned deacc and the latter is all deacc that the system failed to assign.

Table 2. Contingency table. The cases the system got right are called tp = true positives and tn = true negatives and the cases that are wrongly selected are called fp = false positives and the cases that failed to be selected are called false negatives, fn . Deacc is one of the prominence categories. One contingency table is constructed for each prominence category.

	deacc is correct	\neg deacc is correct
deacc was assigned	tp	fp
\neg deacc was assigned	fn	tn

An average of some evaluation measure, e.g. precision and recall, is calculated for all categories from the three contingency tables. There are two methods for averaging the chosen evaluation measure: *micro-averaging* and *macro-averaging*. Micro-averaging is calculated through computing some evaluation measure, e.g. precision, over all categories. When calculating the micro-average precision, the figures of the true positives from the three contingency tables are summed and divided by the sum

of all true positives and all false positives from the three contingency tables according to the precision formula, viz.:

$$\text{Precision} = \text{tp} / (\text{tp} + \text{fp})$$

$$\text{Micro-average precision} = \text{tp}_{(\text{deacc} + \text{acc} + \text{prom})} / (\text{tp}_{(\text{deacc} + \text{acc} + \text{prom})} + \text{fp}_{(\text{deacc} + \text{acc} + \text{prom})})$$

In Figure 5 an example of a prominence labeled sentence, the correct vs. assigned labels, and the contingency tables made from the sentence are given.

Ska	han	dö	innan	den	tusende	strängen	brister?		
(Will he	die	before	the	thousandth	string	breaks?)			
acc	deacc	acc	deacc	eacc	acc	acc	prom	correct	
deacc	acc	acc	deacc	deacc	acc	prom	prom	assigned	

	deacc is correct	¬ deacc is correct
deacc was assigned	2	1
¬ deacc was assigned	1	4
	acc is correct	¬ acc is correct
acc was assigned	2	1
¬ acc was assigned	2	3
	prom is correct	¬ prom is correct
prom was assigned	1	1
¬ prom was assigned	0	6

Figure 5. A sentence with the correct vs. assigned prominence labels is shown at the top of the figure. Below the sentence three contingency tables for the three prominence categories of the sentence are given.

The sentence in Figure 5 gives the micro-average precision of 62.5%, viz.:

$$\text{Micro-average precision} = \text{tp}_{(\text{deacc} + \text{acc} + \text{prom})} / (\text{tp}_{(\text{deacc} + \text{acc} + \text{prom})} + \text{fp}_{(\text{deacc} + \text{acc} + \text{prom})})$$

$$= (2 + 2 + 1) / ((2 + 2 + 1) + (1 + 1 + 1)) = 0.625$$

This gives equal weight to each object, e.g. word. Macro-averaging is computed by averaging over all binary classification decisions, which gives equal weight to each

category (Manning & Schütze, 1999:577). The macro-average precision of the sentence in Figure 5 is hence 61.1% and calculated like this:

$$\text{Macro-average precision} = (\text{tp}_{\text{deacc}} / (\text{tp}_{\text{deacc}} + \text{fp}_{\text{deacc}}) + \text{tp}_{\text{acc}} / (\text{tp}_{\text{acc}} + \text{fp}_{\text{acc}}) + \text{tp}_{\text{prom}} / (\text{tp}_{\text{prom}} + \text{fp}_{\text{prom}})) / 3 = 2 / (2 + 1) + 2 / (2 + 1) + 1 / (1 + 1) \approx 0.667 + 0.667 + 0.5 \approx 0.611$$

In this work only macro-averaging is used to calculate the precision and recall rates. The macro-average is the relevant calculation method because the results of the classification regarding the result of each category are the goal of this thesis, not how the classification results disregarding the categories are. The precision and recall measures are of interest because they reflect the missing target items and the erroneously selected items respectively whereas the accuracy measure does not.

5 Solutions and results

5.1 The prosodic feature vectors to CART

There are two main PFV structures, one for the classification of prominence and one for the prosodic boundaries. These PFV's are tested with and without the prosodic boundary labels and the prominence labels as one of the predictors in the PFV for the classification of prominence and prosodic boundary respectively. The motivation for these additional features being that the probability of a word having some degree of prominence is higher if there is a prosodic boundary before (due to its novelty) or after the word (a default position for prominence) than if there is no boundary (Bruce & Granström, 1993; Fant, Kruckenberg and Liljencrants, 2000:19). This tendency applies to prosodic phrase breaks as well. Classification with PFV's excluding the prosodic labels are done because one cannot assume that speech data with manually labeled prominence or phrase breaks exist. If the results of this classification are satisfying then the only information that is needed as input to CART is taken from the automatically segmented speech databases.

In addition to the prosodic labels, the prosodic feature vectors (PFV) contain mainly duration information, but also features for pause information are a part of the PFV. The duration information is taken from the stressed syllable, the word, the final syllable of the word, the rhyme of the wordfinal syllable, and the difference between the onset and rhyme of the wordfinal syllable. In the PFV for the classification of prominence the pause predictors consist of flags indicating if there is a pause before and after the word. The features for the PFV for prominence labeling are thus:

1. Prosodic phrase break label (*Break Predictor*) (only in one of the PFV's for prominence labeling).
2. Normalized duration for the stressed syllable (*Stress Syll Dur*).
3. Normalized duration for the word (*Word Dur*).
4. A flag indicating if there is a pause before the word or not (*Pause B*).
5. A flag indicating if there is a pause after the word or not (*Pause A*).
6. Normalized duration for wordfinal syllable (*Wordfinal Syll Dur*).
7. Normalized duration for wordfinal rhyme (*Wordfinal Rhyme Dur*).
8. Normalized duration for the difference between the wordfinal onset and rhyme (*Diff Wordfinal Onset Rhyme*).

In addition, the PFV for the classification of prosodic phrase breaks also contains absolute duration of the pause before (*Pause Dur B*) and after the word (*Pause Dur A*) and of course the first feature in the list above is exchanged with prominence labels (*Prom Predictor*).

The normalized duration of the word and the stressed syllable is a measure of the grade of prominence. Experiments show that most of the lengthening in focal accent occurs in the stressed syllable (see 2.2.1). But because these experiments are only made on non-compound words *Word Dur* is a safety feature if this would not hold for compound words and words with a weaker prominence degree, i.e. accentuation (e.g. Nöth et al., 2000:522). The remaining three duration features are chosen to distinguish the words with prominence from the words with final lengthening without prominence. These features are only applied to words with more than one syllable because the features *Stress Syll Dur* and *Word Dur* cover words with only one syllable. *Wordfinal Syll Dur* is thought to capture both the final lengthening phenomenon and prominence where the word with both final lengthening and prominence has the highest value. The *Wordfinal Rhyme Dur* should differentiate between words with and without final lengthening (e.g. Shriberg et al., 2000:132; Whightman & Ostendorf, 1994:471). The difference between the wordfinal onset and rhyme reflects the observations that if only the rhyme is lengthened it is a question of final lengthening without prominence (e.g. Whightman & Ostendorf, 1994:471). The pause features are included due to the fact that pauses often signal major phrase boundaries. (Nöth et al., 2000:523)

5.2 Program development

The process of the automatic labeling, from extracting information from the manually labeled speech databases to the automatic labeling of the speech databases, is divided into a number of smaller processes. See Figure 6 for the main input and output structures with corresponding programs executed for the automatic labeling process. First the features with mainly duration information are extracted for the training and classification process to form the PFV's. The feature extraction programs (A in Figure 6) constitute the main implementations in this work and are described in the next section (5.3). Then the CART program Wagon (B in Figure 6) is run on the training and testing data (as described in 4.1.1) with the PFV's and a description file for the PFV in question. The trained trees from Wagon are further parsed so that they are usable in the annotation program, because the output from Wagon does not contain the predicted labels for each word, which is needed in the labeling. Two already existing programs, at Babel-Infovox AB, are executed for this purpose. The first program, the CART parser, (C in Figure 6) analyzes the trees from Wagon and rewrites them to XML trees. The second program, the CART classifier, (D in Figure 6) analyzes the XML tree and writes the predicted labels for each word into a classification file. The classification file is then input to the annotation program, (E in Figure 6) which writes the predicted labels to each word in the XML database (see section 5.4 for details on the annotation program). The automatically labeled speech databases are finally evaluated.

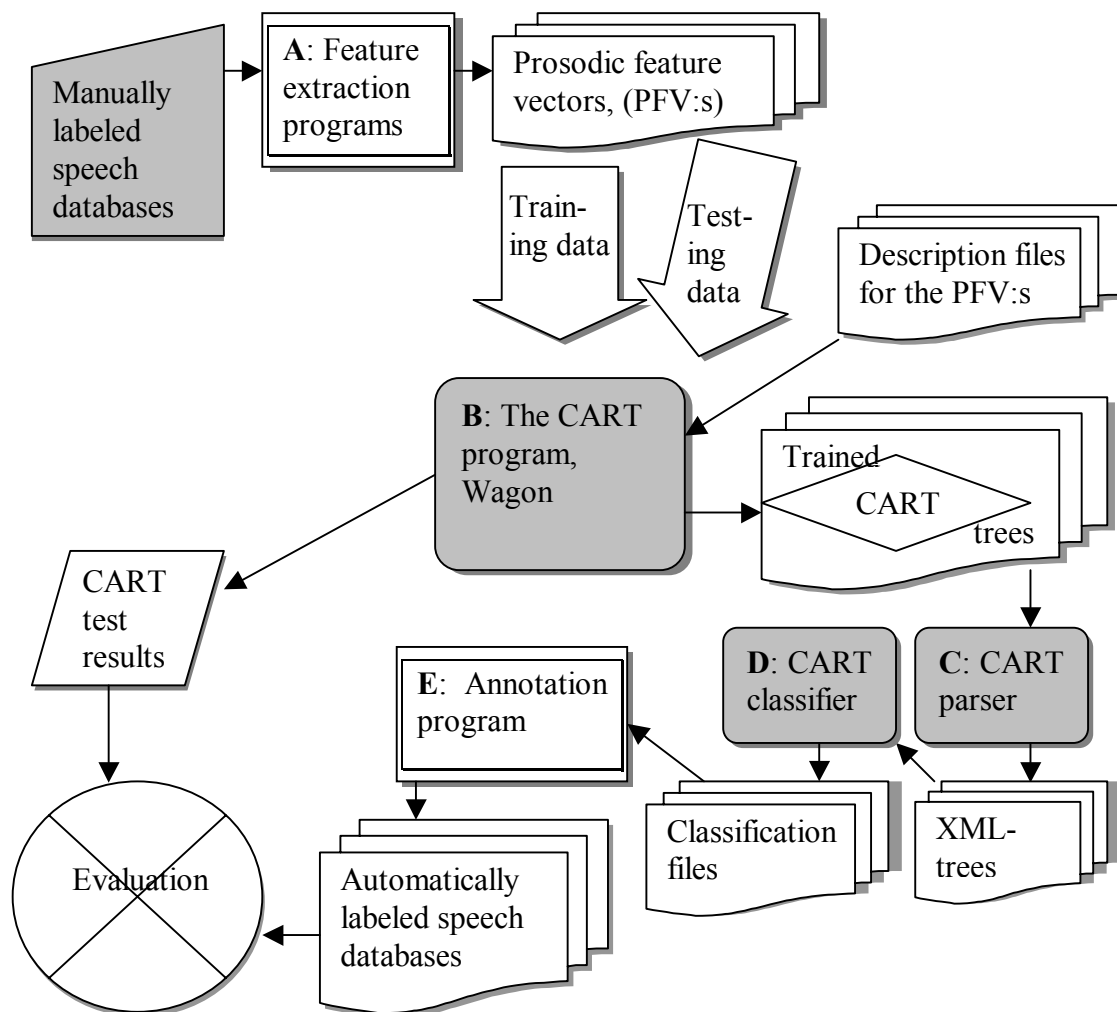


Figure 6. The process of the automatic labeling, from the manually labeled speech databases to the automatically labeled speech databases. The gray boxes are already existing programs, i.e. programs not made in this thesis.

5.2.1 Algorithms and implementations

Before the actual feature extraction (A in Figure 6), the randomization of the sentences in the training data is done (A in Figure 7), because it is only the original XML databases that contain the segmentation and order of the sentences that is needed in order to randomize the sentences. Then the duration of each segment is extracted from the training data in order to calculate the mean and standard deviations for each phoneme (B in Figure 7), which are used in the normalization calculations for

the features in the PFV's. Thereafter the actual feature calculations can be done (C in Figure 7). Figure 7 shows how the main feature extraction processes are conducted.

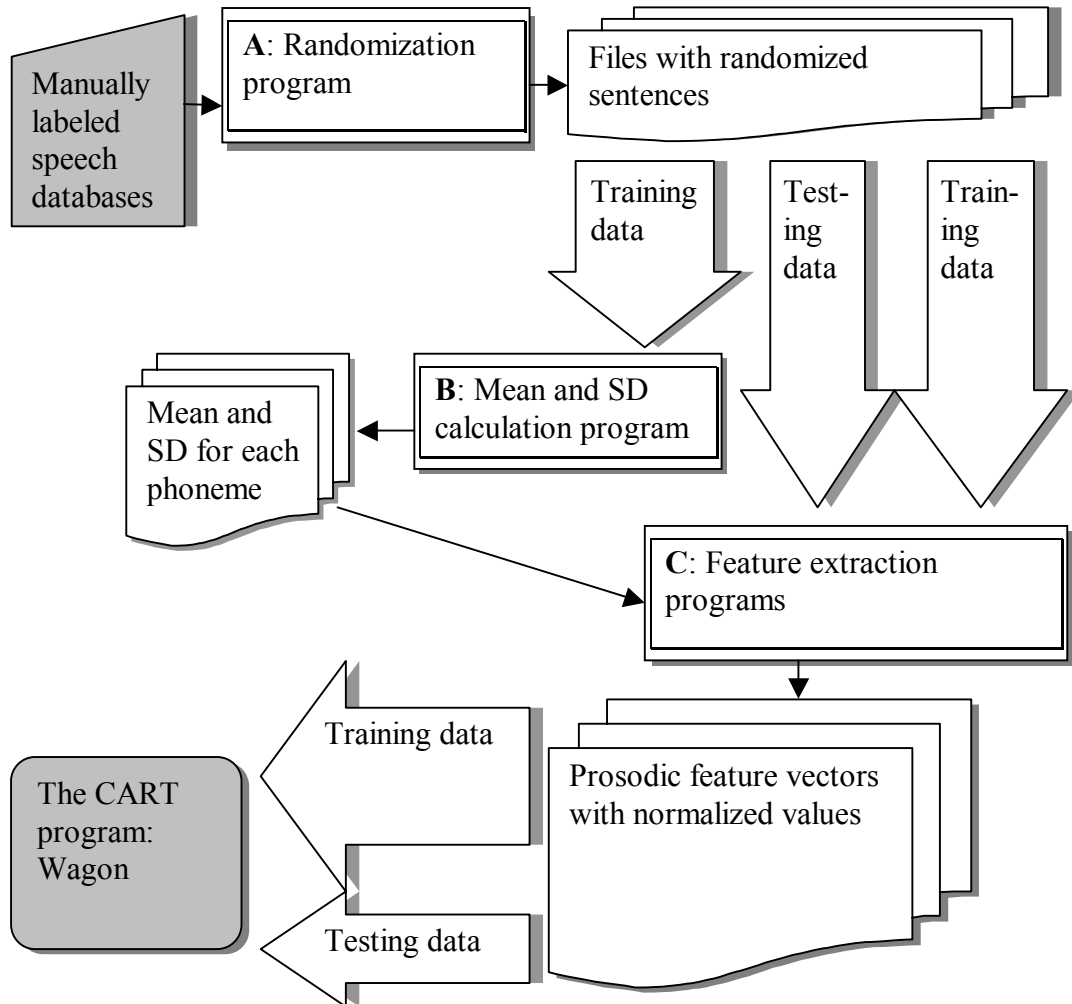


Figure 7. The feature extraction processes. The gray boxes are already existing programs, i.e. programs not made in this thesis.

All programs developed in this work are implemented in the computer language Perl. The main reason for this is that it handles string processes well, which is one of the main tasks of the programs. In Appendix 2 the code for the programs is presented. When there are a number of programs with only slightly different code only one of them is presented in the appendix, e.g. the programs for the data extraction.

The sentence randomization (random.perl). All sentences in the nine speech databases are randomly divided into five equal parts, which will be used in the 5-folded cross-validation in the training and testing of the prosodic classification. First all sentences in the speech databases, which are segmented by the tag <sentence>, are searched and put into an array. Then the randomization is done for each sentence in the list, with the Perl command ‘rand’, which randomly gives a number between 0.0 and 1.0. This range is divided into five equal parts which gives a condition for the division of the sentences, i.e. if ‘rand’ has given a number between the range for the first group, $0.0 > < 0.2$, then the sentence is put into outfile 1, and so forth.

The mean (M) and standard deviation (SD) calculations (mean_stand.perl). The mean and standard deviation for each phoneme are calculated by first collecting all phones together with their duration in milliseconds which are segmented in the databases. The values of the phones and duration become substrings in a hash table. Then the phonemes in the hash table are sorted by the ‘sort keys’ command so that there is only one /a/, for example, in the phoneme substring and all durations for /a/ is summed up and divided by the number of phones corresponding to /a/. When the mean value for each phoneme is given, the SD can be calculated by using the same hash table to get the milliseconds for each phone again. The output of this program is a file containing a list of the phoneme names, the mean value and the SD for the phoneme.

The data extraction (prom.perl, prom_w_prosBreak.perl, pros-Break.perl and prosBreak_w_prom.perl) . There are four programs in order to do the four different PFV’s: prom.perl that calculates the feature values for the PFV used in the classification of prominence; prom_w_prosBreak is extracting the prosodic phrase break labels and add it to the other features used for prominence classification to build the second PFV. The third and fourth PFV, which are used in the classification of prosodic phrase breaks, are made by prosBreak.perl and prosBreak_w_prom.perl respectively with the difference that the prominence labels are included in the PFV

from the latter program. These programs are very similar since the features for the two prosodic phenomena are almost the same. Therefore an algorithm is briefly presented in the list below for `prom_w_prosBreak.perl` with the most important solutions for the implementation included. Above this, the algorithm for `pros-Break.perl` and `pros-Break_w_prom.perl` also includes extraction of the pause duration before and after the word (Pause Dur B and Pause Dur A).

- 1. Fetch the mean and standard deviation values from the outputfile of `mean_stand.perl`.** The mean and SD values are put into a hash table with substrings for the phoneme, the mean duration and SD of the phoneme so the information is easily accessible.
- 2. Make a list of all words in the training data.** The data from the training files are divided into an array of tokens, `@token_list`, segmented by the tag `<token>` in the database, with the command `'split'`.
- 3. For each word, note the prosodic phrase break label.** A `'foreach'`-loop checks each item in the array `@token_list` if `type="word"`, i.e. if the item is a word, if not, next item in the array is checked by the command `'next'`. In the same loop the word is checked if it has a prominence label. If there is a prominence label it is written into the PFV.
- 4. For each word, make a list of all phonemes in the word.** An array, `@phoneme_list_word`, is created with the command `'split'`.
- 5. For each word, make a list of its syllables.** This is solved in the same way as for the words. The syllables are segmented with the tag `<syllable>`.
- 6. For each syllable in each word, make lists of the phoneme segments.** This is solved in the same way as for words and syllables. The phones are segmented like this: `<phoneme symbol="a" duration="70">`.
- 7. For each syllable in each word, note the stress degree.** If `<accent>` equals 1 or 2 it is stressed, if it has a 0 or a 4 it is unstressed.
- 8. For each stressed syllable, calculate the z-score and write it to the PFV.** The z-scores of all segments in the syllable are summed up and divided by the number of segments in the syllable.
- 9. Count every syllable.**

10. **If the word has more than one syllable, then calculate the z-score for the last syllable for the feature Wordfinal Syll Dur and write it to the PFV.** The last syllable is the last item in the syllable array.
11. **Calculate the z-score for the rhyme in the last syllable and write it to the PFV.** This is calculated by calculating the z-score for the vowel and the consonant/s in the rhyme in the last syllable. The vowels and consonants are detected by looking up phoneme/s in a vowel hash table and a consonant hash table.
12. **Calculate the difference between the onset and rhyme duration and calculate the z-score of the difference and write it to the PFV.** The onset is detected by finding the vowel in the syllable, which belongs to the rhyme, and then calculating the z-scores for each consonant before the vowel. If, for example, it is the third item in the array of phonemes in the syllable that is the vowel, then item number one and two belong to the onset, i.e. \$phoneme_list[1] and \$phoneme_list[2]. The program handles onsets with one, two or three consonants.
13. **Count every word.**
14. **For each word, check if there is a pause before the word. Write the presence or absence of a preceding pause into the PFV.** This is done in the 'foreach'-loop of @token_list. The word that is processed has the same item number as the \$count_word, and the item before is hence \$count_word - 1. The item before the word that is processed is checked if it has a pause sign or not. This information is written into the PFV.
15. **For each word, check if there is a pause after the word and write the information into the PFV.** The same solution as for the preceding pause holds for this problem too.
16. **For each word, calculate the z-score and write it to the PFV.** This is done by calculating the z-score for each phone in the word, i.e. in the array @phoneme_list_word.

The annotation (annotate_prom.perl, annotate_prosBreak.perl). These programs write the predicted labels from the classification files into the XML databases with manually labeled prominence or prosodic phrase breaks. First the predicted labels are read from the classification file and put into an array, @prom_label. Then the XML database that will be annotated is read and checked if type="word" and further if

there is a prominence tag. If so, the predicted label for that word is printed next to the manually labeled prominence so that they can be compared.

The evaluation (eval_prom.perl, eval_prosBreak.perl). These programs count the number of correct and false annotations and put the numbers into a contingency table. This is done by checking if the type is a word and what the manually labeled and automatically labeled prominence or phrase break are. For each word the kind of labels are counted, e.g.:

```
if ( /prominence="deacc:deacc/ && ( $isWord eq 'yes' ) ){
    $count_deacc_deacc ++;
}
elseif ( /prominence="deacc:acc/ && ( $isWord eq 'yes' ) ){
    $count_deacc_acc ++;
}
etc.
```

The information in the contingency tables from the evaluation programs is then used in the calculations of the measures precision and recall. These results are then compared with the evaluated results from CART, which also outputs contingency tables.

5.3 Evaluation and discussion

5.3.1 CART results and annotation results

Before the training and testing of CART a few corrections and modifications were done on the speech databases due to errors found when testing the machine learning technique. Sentences lacking segment duration annotations were removed. Furthermore, incorrect prominence labels were corrected.

The measures calculated for each prominence and boundary class in the contingency tables of the five test parts and the contingency table for the labeling

result, are recall and precision. The mean value of the classification result for each contingency table for these measures is calculated with macro-averaging. Finally, a mean and a standard deviation value for all five mean values of the classification contingency tables are calculated (see Table 3 and 6).

Prominence. The results show that duration information and other segmental information are highly usable in the classification of prominence. As expected, the classification of prominence with the Break Predictor produced better results than without (see Table 3). The recall mean value of all test parts when using the Break Predictor in the training is 75.98% versus 69.29% when the tree was trained without it. The precision rate is 69.62% without the Break Predictor and 77.39% with the predictor.

The tests with speaker dependent speech data, i.e. the speech databases read by one person, (here the speech databases read by the male) gave a little lower recall, 75.42% and precision, 77.24% than classification with the speaker independent speech data (see Table 3). This is probably because the speaker dependent speech data is smaller, around half the training data as for the speaker independent speech data.

The trees produced with the Break Predictor were chosen for the actual labeling of the speech databases since these gave the best results compared to the trees produced without the Break Predictor. The stopping of the tree gave slightly better results; recall rate reached 76.22% and the precision rate reached 78.71% (see Table 3). The reason why the improvement of the stopping is not greater is probably due to the method of reaching the best size for a tree. The measure, given by Wagon, and used as a measure of how good result the stopping of the tree is, are micro-averaged recall or precision, and not macro-averaged recall or precision which are the measures used in this thesis. An improvement of the recall or precision rate, calculated by micro-averaging, does not necessarily imply an improvement of the recall and precision rate calculated by macro-averaging.

Table 3. Cross-validation results of CART classification of prominence. The mean and standard deviation values for the mean value (macro-averaged) of the three prominence classes: deacc, acc and prom are calculated. ‘Prom’ is classification with duration and pause features, ‘Prom + Break Predictor’ is classification with the prosodic break label as one of the predictors and the row ‘Prom + Break Predictor + stop’ contains mean values of the trees classified with the same predictors as ‘Prom + Break Predictor’ but with the stopping criteria.

	Recall		Precision	
	M	SD	M	SD
Prom	69.29%	0.87%	69.62%	1.01%
Prom + Break Predictor	75.98%	4.73%	77.39%	1.02%
Prom + Break Predictor + stop	76.22%	0.15%	78.71%	0.73%
Speak. dep. Prom + Break Predictor + stop	75.42%	1.20%	77.24%	0.27%

In Table 4 the differences between the three categories are presented. The mean values of recall and precision from the classification with only duration features are higher for the category deacc, 82.93% and 75.99% respectively, than for acc and prom; 62.27% and 62.67% recall and 65.12% and 67.77% precision. This relationship also holds for the recall rate when the Break Predictor was incorporated into the PFV whereas the precision rate for prom is the highest, 89.22%, compared to deacc and acc: 76.84% and 72.57%. The increased recall and precision rates, from using Break Predictor seen in Table 3 above, are though almost exclusively due to the results of acc and prom which have increased with from 10 to almost 20 percentage points for both the recall and precision rates, (see Table 4).

Table 4. The mean (macro-averaged) and standard deviation values for each prominence class. ‘Prom’ is classification with only duration and pause features and ‘Prom + Break Predictor + stop’ is classification with the prosodic phrase break label as one of the predictors and a stopping criterion.

	Recall		Precision	
	M	SD	M	SD
Prom, deacc	82.93%	1.06%	75.99%	0.33%
Prom, acc	62.27%	2.61%	65.12%	1.64%
Prom, prom	62.67%	1.06%	67.77%	3.59%
Prom + Break Predictor + stop, deacc	84.59%	1.01%	76.84%	3.50%
Prom + Break Predictor + stop, acc	72.02%	2.45%	72.57%	2.13%
Prom + Break Predictor + stop, prom	72.05%	1.80%	89.22%	0.23%

The best-sized tree, with a stopping number of 35, was then trained and tested on all subsets, which gave a small improvement (see Table 5). The improvement is probably due to the larger amount of training data. The improvement is however not very big, which is probably due to that the best size of the tree is just an average of all stopping numbers for the subset trees, which does not mean that it is the best number for the final tree. The best-sized tree was used in the labeling of the speech databases. The results of the labeling, which are shown in Table 5, are merely as good as the CART results. The small difference can be due to the fact that the trees produced for the labeling were based on non-randomized data in order to match the order of words in the classification file with the order in the speech databases.

Table 5. Prominence classification and labeling. In the first row of the table the mean values of recall and precision of all stopped trees are shown. In the second row the results of running CART with the best-sized tree, (mean stopping number: 35) on all test data are shown, and in the last row of the table, the results of the annotation with the same tree as training data are presented.

	Recall	Precision
CART results, mean value of all trees: Prom + Break Predictor + stop (macro)	76.22%	78.71%
CART results, the best-sized tree: Prom + Break Predictor + stop (macro)	76.96%	79.85%
Annotation results: Prom+ Break Predictor + stop (macro)	76.67%	79.04%

Boundaries. The classification of prosodic boundaries with only segmental information did not give as good average results as the classification of prominence. Only a mean value of 49.68% recall and 51.61% precision were reached when the boundaries were classified with only duration features (see Table 6). Like the results of the prominence classification, an improvement to a recall rate of 57.98% and a precision rate of 54.45% occurred when using the Prom Predictor in the PFV. The speaker dependent classification resulted in slightly lower recall and precision rates for boundary classification than the prominence classification; 58.28% and 53.28% (see Table 6).

Like the choice of trees for the actual labeling of prominence, the trees produced with the Prom Predictor were chosen for the labeling of boundaries since these also gave the best results. The results of the stopped trees were only improved by at most 0.35 percentage points.

Table 6. Cross-validation results of CART classification of boundaries. The mean and standard deviation values for the mean value (macro-averaged) of the three boundary classes: NB, SB and BB are calculated. 'ProsB' is classification with duration and pause features, 'ProsB + Prom Predictor' is classification with the prominence label as one of the predictors and the row 'ProsB + Prom Predictor + stop' contains mean values of the trees classified with the same predictors as 'ProsB + Prom Predictor' but with the stopping criteria. Last row is classification with speaker dependent speech data.

	Recall		Precision	
	M	SD	M	SD
ProsB	49.68%	0.58%	51.61%	1.17%
ProsB + Prom Predictor	57.98%	0.91%	54.45%	0.32%
ProsB + Prom Predictor + stop	58.33%	0.51%	54.50%	0.58%
Speak. dep. ProsB + Prom Predictor + stop	58.28%	1.39%	53.28%	0.76%

The results of the different categories are presented in Table 7. The low figures of the average results, over the categories, of the classification of boundaries are due to the failure of the classification of Small Breaks. The classification failed to classify any SB correct. This could depend on the fact that the training data of SB was very small, only around 130 SB:s were labeled manually in each subset, compared to around 3600 NB:s and 670 BB:s. Another reason could be that duration features are not descriptive enough for this phenomenon. Perhaps F_0 features, voice quality features and energy features would be better descriptors. The results of classifying BB are better than SB, the recall is 53.11% and the precision 64.98%. The words without a prosodic phrase break, NB, have the best recall, 95.92% and precision rates, 89.84%. This is probably due to the high number of training examples (around 3600 NB:s).

In Table 7 the small decrease of the recall rate for NB, when classifying with a stopped tree, from 95.92% to 95.04%, are probably the result of the method for finding the best size of a tree. Otherwise, the stopping of the trees has resulted in

small improvements of the classification. The precision rate of NB is improved from 89.84% to 94.46%. The recall rate for BB increased from 53.11% to 79.95% and the precision rate from 64.46% to 69.04%.

Table 7. The mean and standard deviation values for each boundary class. ‘ProsB’ is classification with only duration and pause features and ‘ProsB + Prom Predictor + stop’ is classification with the prominence label as one of the predictors and a stopping criterion.

	Recall		Precision	
	M	SD	M	SD
ProsB, NB	95.92%	0.80%	89.84%	0.58%
ProsB, SB	0.00%	0.00%	0.00%	0.00%
ProsB, BB	53.11%	2.41%	64.98%	4.06%
ProsB + Prom Predictor + stop, NB	95.04%	1.20%	94.46%	1.08%
ProsB + Prom Predictor + stop, SB	0.00%	0.00%	0.00%	0.00%
ProsB + Prom Predictor + stop, BB	79.95%	1.31%	69.04%	1.14%

The results of the best-sized tree are at most only around one percentage point better, which can be seen in Table 8. The reason for the improvement is probably the same as for the prominence classification: the larger amount of training data. The reason for why the improvement is not greater could be the same as for the prominence classification; the stopping number of the tree is an average of the stopping numbers for the training trees and not adjusted to the new unseen test data. It could also be a result of the stopping number, 50, which gives the same result as without any stopping of the tree. The results of the labeling are shown in Table 8. They are the same as the CART results. As previously mentioned, this is due to the fact that the stopping number is 50.

Table 8. Boundary classification and labeling. In the first row of the table the mean values of recall and precision of all stopped trees are shown. In the second row the results of running CART with the best-sized tree, (mean stopping number: 50) on all test data are shown, and in the last row of the table, the results of the annotation with the same tree as training data are presented.

	Recall	Precision
CART results, mean value of all trees: ProsB + Prom Predictor + stop	58.33%	54.50%
CART results, the best-sized tree: ProsB + Prom Predictor + stop (macro)	59.54%	54.80%
Annotation results: ProsB_w_prom_stop (macro)	59.54%	54.80%

Features chosen by CART. When classifying, CART chooses, for each node in the tree, a feature to formulate a question of, which gives the best division of the training data. The feature in the question in the root node is hence the most important feature because it gives the highest information gain.

Prominence features. The features in the classification of prominence without the Break Predictor are, in decreasing order, the most important:

1. Word Dur
2. Wordfinal Rhyme Dur
3. Stress Syll Dur
4. Diff Wordfinal Onset Rhyme
5. Wordfinal Syll Dur

The pause features, Pause B and Pause A were not used in the classification, which is remarkable because pause information should be useful providing the probabilities of focused words in the end and beginning of sentences (see section 5.1). The most important features in the classification with the Break Predictor are:

1. Break Predictor
2. Stress Syll Dur
3. Diff Wordfinal Onset Rhyme
4. Word Dur
5. Wordfinal Rhyme Dur

The Word and Wordfinal Rhyme Dur are replaced by the Break Predictor, which is reasonable because they can describe the same thing, namely the lengthening of the word when it is focused. Besides the pause features, the Wordfinal Syll Dur is not used in this classification.

Prosodic boundary features. The features excluded by CART in the classification of prominence are the same as the features excluded in the classification of prosodic boundaries, i.e. the pause features Pause B and Pause A and the Wordfinal Syll Dur. In addition the features Pause Dur B and Pause Dur A are not used either. This is extraordinary as these features contain information about the prosodic breaks' duration. A reasonable explanation to the fact that CART excludes the pause features could be that the Small Breaks do not have any pauses at all, but are signaled through other acoustic means and could not therefore be compared to the Big Breaks with respect to the pauses. The most important features when classifying without the Prom Predictor are:

1. Wordfinal Rhyme Dur
2. Word Dur
3. Stress Syll Dur
4. Diff Wordfinal Onset Rhyme

In the classification with the Prom Predictor the dropped features are the same as in the classification without the Prom Predictor. The most important features are:

1. Prom Predictor
2. Wordfinal Rhyme Dur
3. Word Dur
4. Diff Wordfinal Onset Rhyme
5. Stress Syll Dur

The order of the predictors in the classifications of boundary is very similar. The only difference is the order of the two last, number four and five.

When CART chooses the best feature for each node it do not test what feature is the best at that node regarding the result of the whole tree but only the result for the

division at that node. Accordingly, one can not say that the order of the features presented above are the absolute best, one can only say that it is an order that do not give the worst results. The order of the features and what features are the best should hence be further investigated by experimenting with different combinations.

Comparison with previous work. A comparison of the results of this work and the results of other automatic prosodic classifications does not say so much about the results of this work. This is partly because the measures presented by other researchers are not relevant or even the same as those presented in this work and partly because the classification has been conducted with different features, categories, training data, and machine learning technique. However, a small comparison is made of the results of this thesis and the results of two other important works; the work of Wightman & Ostendorf (1994) and the Verbmobil project (Nöth et al., 2000) in order to give a picture of the quality of the results of this thesis.

Prominence classification. Wightman & Ostendorf classified intonation markers using features describing among other things duration, F_0 and energy (1994). The English training data contained 2140 words spoken by three professional radio news announcers. The intonation marker categories were unmarked syllable (S), accented syllable (P), boundary tone (BT) and syllables with both an accent and a boundary tone (P-BT). The average recall and precision rates for the intonation marker classification excluding the class of boundary tones were 79% and 75% respectively compared to the best results of this work; 77% and 80% for recall and precision (see Table 9) (Wightman & Ostendorf, 1994).

In the Verbmobil project both prominence and prosodic boundaries were given binary classification, i.e. $\neg A/A$ and $\neg B/B$ respectively, using features like duration, F_0 and energy, with a corpus of 17 h of speech containing 514 words. The average recall rate when using all features was 82% compared to 77% in this thesis. When using only duration features the average recall rate is 75%, which is more comparable with the results of the classification without the Break Predictor (69%), although this

classification is not optimized through stopping the tree (see Table 9) (Nöth et al., 2000).

Table 9. Prominence classification results in this thesis compared to the results of Wightman & Ostendorf (1994) and the results of the Verbmobil project (2000). The result of this work is the result of running CART with the best-sized tree with the Break Predictor included in the PFV and in the second row without the Break Predictor in the PFV to be compared to the last row in the table. The third row shows the average of all intonation markers but boundary tone of the work of Wightman & Ostendorf. The fourth row shows the result prominence classification of Verbmobil using all features and the last row shows the classification result of using only duration information. All figures are macro-averaged recall and precision.

	Recall	Precision
Results of this work: deacc, acc & prom + Break Predictor	77%	80%
Results of this work: deacc, acc & prom	69%	-
Wightman & Ostendorf: S, P & P-BT	79%	75%
Verbmobil: ¬A/A	82%	-
Verbmobil: only duration information, ¬A/A	75%	-

Although it is difficult to compare these results it is remarkable that the classification of prominence in this work, using mainly duration cues, gives almost as high classification rates as the works compared with in this section (the work of Wightman & Ostendorf (1994) and the Verbmobil project (Nöth et al., 2000)), which have used other important acoustic features besides duration cues, (see Table 9).

Prosodic boundary classification. Like the classification of intonation markers Wightman & Ostendorf classified prosodic breaks using among other features, different duration, F_0 and energy features. The English training data were the same as for the intonation markers. A 7-level break-index was used for the labeling of the prosodic boundaries. The average recall and precision of the 7-level prosodic boundary classification by Wightman & Ostendorf is 41% and 42% respectively (1994:478) compared to an average recall and precision rate of 60% and 55% respectively in this work (see Table 10).

The prosodic boundaries were like the prominence also binary classified in the Verbmobil project, (¬B/B) using features like duration, F_0 and energy, with the same

corpus as for the classification of prominence. The average recall and precision rate are 87% and 74%, respectively. When classifying with only duration cues the recall is 78% compared to 50% of this work (although the classification without the Prom Predictor is not optimized) (see Table 10).

Table 10. Prosodic boundary classification results in this thesis compared to the results of Wightman & Ostendorf (1994) and the results of the Verbmobil project (2000). The result of this work is the result of running CART with the best-sized tree with the Prom Predictor included in the PFV and in the second row without the Prom Predictor in the PFV to be compared to the last row in the table. The third row shows the average of the seven break indices of the work of Wightman & Ostendorf. The fourth row shows the result of boundary classification of Verbmobil using all features and the last row shows the classification result of using only duration information. All figures are macro-averaged recall and precision.

	Recall	Precision
Results of this work: NB, SB, BB + Prom Predictor	60%	55%
Results of this work: NB, SB, BB	50%	-
Wightman & Ostendorf: Breaks 0-6	41%	42%
Verbmobil: ¬B/B	87%	74%
Verbmobil: only duration information, ¬B/B	78%	-

An interesting remark about the classification of prosodic boundaries is that the fewer categories used in the classification the better the results seem to be. In addition, the huge corpus used in Verbmobil, surely improved the results which are the best in this comparison.

Even though these comparisons above are problematic, a conclusion that can be drawn from the results in notion of the results in this work, however, is that one can come a long way with only duration features in the classification of prominence and even longer with additional segmental features. Duration features are helpful in the classification of prosodic boundaries as well, but not as good as for prominence.

6 Conclusions

In this thesis it has been shown that duration cues are indeed important for distinguishing the different levels of prominence. The classification of prominence reached a mean recall of 69% and a mean precision of 70% for speaker independent recognition. When using the hand labeled phrase breaks as additional cues to prominence the recall and precision increased to 76% and 78%, respectively. The labeling of phrase breaks needs other cues than duration features and pause information to be able to distinguish between the small and big breaks.

The automatic prominence labeler can be used at this moment, perhaps more as a prelabeler, than instead of the manual labeling work. In order to improve the results of the automatic labeling of prominence in Swedish, one would probably need to incorporate other kinds of phonetic information, such as F_0 and energy features.

6.1 Further developments

There is work that could be done to improve the classification of prominence and prosodic boundaries. First, the duration features could be further investigated by testing different combinations of the already existing duration features in the training of CART, but also new duration features could be explored. Moreover, other segmental information could easily be implemented, such as part-of-speech.

Refinements of the training procedure could also be done to further improve the results, with for example larger training data and a better stopping procedure of trees. The decision of the best stopping number of a tree should be based on the same measure as in the evaluation. Explorations of other machine learning techniques, such as Multi Layer Perceptrons, could also be done, in order to compare which technique produces the best results.

Furthermore, using additional phonetic features, like F_0 and energy features, would probably lead to better classification results. Not only phonetic features should be explored however, the use of other linguistic information, such as syntax and semantics, would be interesting to incorporate in the system as well.

References

- Batliner, A., E. Nöth, J. Buckow, R. Huber, V. Warnke and H. Niemann, (2001a): Duration Features in Prosodic Classification: Why Normalization Comes Second, and what they Really Encode. *Proc. ISCA Tutorial and Research Workshop on Speech recognition and understanding*, Red Bank, New Jersey. 23-28.
- Batliner, A., B. Möbius, G. Möhler, A. Schweitzer and E. Nöth, (2001b): Prosodic models, automatic speech understanding, and speech synthesis: towards the common ground. *Eurospeech 2001 Scandinavia*, chairman: P. Dalsgaard. Aalborg.
- Beckman, M. E. & G. Ayers Elam, (1997): *Guidelines for ToBI labelling, version 3*. The Ohio State University Research Foundation. Retrieved 2002-10-01, from: http://www.ling.ohio-state.edu/research/phonetics/E_ToBI/.
- Breiman, L., (1984): *Classification and regression trees*. California: Wadsworth, Inc.
- Bruce, G., (1998): *Allmän och svensk prosodi*. Praktisk lingvistik 16. Lund: Dept. of Linguistics, Univ. of Lund.
- Bruce, G. & B. Granström, (1993): Prosodic Modelling in Swedish speech synthesis. *Speech Communication* 13:63-74.
- Bruce, G., B. Granström & D. House, (1992): Prosodic phrasing in Swedish speech synthesis. *Talking Machines: Theories, Models and Designs*, ed. by G. Bailly, C. Benoit. Amsterdam. 307-321.
- Campbell, N., (1993): Automatic detection of prosodic boundaries in speech in *Speech Communication* 13:343-354.
- Cutugno, F., L. D'Anna, M. Petrillo and E. Zovato, (2002): APA: Towards an automatic tool for prosodic analysis. *Proc. of the 1st Int. Conf. on Speech Prosody*, chairman: Daniel Hirst. Aix-en-Provence.
- Elert, C-C., (1991): *Allmän och svensk fonetik*. Stockholm: Norstedts Förlag AB.
- Fant, G., A. Kruckenberg and L. Nord, (1990): Acoustic correlates of rhythmical structures in text reading. *Nordic Prosody V*, ed. by K. Wiik & I. Raimo. Turku: Phonetics Dept., Univ. of Turku. 70-86.

- Fant, G., A. Kruckenberg & L. Nord, (1997): *Some studies of accent and juncture in Swedish*. Phonum 4:157-160. Umeå: Phonetics dept., Univ. of Umeå.
- Fant, G., A. Kruckenberg, J. Liljencrants and A. Botinis, (2001): Prominence correlates. A study of Swedish. *Eurospeech 2001 Scandinavia*, chairman: P. Dalsgaard. Aalborg.
- Fant, G., A. Kruckenberg and J. Liljencrants, (2000): Acoustic-phonetic Analysis of Prominence in Swedish. *Intonation, Analysis, Modeling and Technology*, ed. by Antonis Botinis. Dordrecht: Kluwer Academic Publishers. 55-86.
- Frank, E., (2003): http://www.mkp.com/books_catalog/weka/teaching_material/ML_part_II.pdf. Univ. of Waikato.
- Greenberg, S., (2002): From here to utility? Melding phonetic insight with speech technology. To appear in *Integrating Phonetic Knowledge with Speech Technology*, ed. by W. Barry and W. Domelen. Dordrecht: Kluwer.
- Gussenhoven, C. & H. Jacobs, (1998): *Understanding phonology*. Understanding language. London: Arnold.
- Heldner, M., (1996): Phonetic correlates of focus accents in Swedish. *TMH-QPSR*, 2:1-4. Stockholm: Department of TMH, KTH.
- Heldner, M., (2001) On the non-linear lengthening of focally accented Swedish words. *Nordic Prosody: Proceedings of the VIIIth Conference*, ed. by W. van Dommelen & T. Fretheim. Frankfurt am Main: Peter Lang. 103-112.
- Heldner, M & E. Strangert, (2001): Temporal effects of focus in Swedish. *Journal of Phonetics* 29:329-361.
- Horne, M., E. Strangert & M. Heldner, (1995): Prosodic boundary strength in Swedish: final lengthening and silent interval duration. *Proceedings ICPhS-95*, ed. by K. Elenius & P. Branderud. Stockholm: Dept. of Speech Communication and Music Acoustics, KTH and Dept. of Linguistics, Stockholm University. 1:170-173.
- House, D. & G. Bruce, (1990): Word and focal accents in Swedish from a recognition perspective. *Nordic Prosody V*, ed. by K. Wiik & I. Raimo. Turku: Phonetics Dept., Univ. of Turku. 156-173.
- Krippendorff, K., (1980): *Content analysis*. Beverly Hills: Sage Publications.

- Llisterri, L., (2003): <http://www.ilc.pi.cnr.it/EAGLES96/spokentx/node31.html>. Univ. of Barcelona.
- Manning, C. D. & H. Schütze, (1999): *Foundations of statistical natural language processing*. Cambridge, Mass.: MIT Press.
- Mitchell, T. M., (1997): *Machine Learning*. New York: McGraw-Hill.
- Mixdorff, H., (2002): Speech Technology, ToBI, and Making Sense of Prosody. *Proc. of the 1st Int. Conf. on Speech Prosody*, chairman: Daniel Hirst. Aix-en-Provence.
- Navas, E., I. Hernáez and N. Ezeiza, (2002): Assigning Phrase Breaks Using CARTs for Basque TTS. *Proc. of the 1st Int. Conf. on Speech Prosody*, chairman: Daniel Hirst. Aix-en-Provence. 527-531.
- Nöth, E., A. Batliner, A. Kießling, R. Kompe and H. Niemann, (2000): Verbmobil: The use of Prosody in the Linguistic Components of a Speech Understanding System. *IEEE Trans. Speech and Audio Processing* vol 8, 5:519-532.
- Pierrehumbert, J. B., (1980): *The Phonology and Phonetics of English Intonation*. Unpublished Ph D Thesis, distributed by Indiana University Linguistics Club 1987, Bloomington, Indiana.
- Portele, T & B. Heuft, (1997): Towards a prominence-based synthesis system. *Speech Communication* 21:61-72.
- Shriberg, E., A. Stolcke, D. Hakkani-Tür and G. Tür, (2000): Prosody-based automatic segmentation of speech into sentences and topics. *Speech Communication* 32:127-154.
- Strangert, E. & M. Heldner, (1994): Prosodic labelling and acoustic data. *Working Papers 43: Papers from the Eighth Swedish Phonetics Conference*, ed. by G. Bruce, D. House & P. Touati. Lund: Dept. of Linguistics and Phonetics, Lund University. 120-123.
- Strangert, E., & M. Heldner, (1998): On the amount and domain of focal lengthening in Swedish. *ICSLP'98 Proceedings* ed. by R. H. Mannell & J. Robert-Ribes. Sidney: ASSTA. 3305-3308.
- Taylor, P., (1998). *Analysis and Synthesis of Intonation using the Tilt Model*. Draft paper. Retrieved 2002-10-01, from: <http://www.cstr.ed.ac.uk/~pault/papers.html/>.

- Taylor, P., R. Caley, A. W. Black and S. King, (2003): http://festvox.org/docs/speech_tools-1.2.0/c16616.htm. Centre for Speech Technology, Univ. of Edinburgh.
- Wells, J. C., (2003): <http://www.phon.ucl.ac.uk/home/sampa/samprosa.htm>
- Wightman, C. W. & M. Ostendorf, (1994): Automatic Labeling of Prosodic Patterns. *IEEE Trans. Speech and Audio Processing*, vol 2, 4:469-481.
- Wightman, C. W., S. Shattuck-Hufnagel, M. Ostendorf and P. J. Price, (1992): Segmental durations in the vicinity of prosodic phrase boundaries. *Journal of Acoustical Society of America*, vol 91, 3:1707-1717.

Appendix 1: Vocabulary index

Vocabulary index

This is a small vocabulary that also works as an index for the terms not explained in the chapters, for the most important phonetic and phonological terms and for other, important terms for this thesis, explained in the chapters. The definitions of the terms are made for this thesis. The page numbers are written ranking the most important page first.

A

Acoustics

The study of sound.

Acoustic correlate 7, 10

The acoustic signal that correlates to a phenomenon of prosody, e.g. prominence or prosodic boundary.

Automatic speech recognition (ASR)

6

A technique that automatically “recognizes” speech, in e.g. phones, syllables and words (compare Speech understanding system).

Automatic speech understanding (ASU) 6, 13

A technique that automatically “understands” speech, i.e. analyzes the grammatical and semantic components of the speech. Speech recognition (see Speech recognition system) is usually included in a speech understanding system.

B

Boundary 7, 12, 27, 34, 36, 39

See Prosodic boundary.

Boundary tone 14, 16, 40

Boundary tones are the tones in a sentence conveying whether it is a statement or a question.

C

CART 18, 19-22, 17, 8, 3,

27 f, 33 f, 36, 38 f, 43

Classification And Regression Trees, a decision tree algorithm within the machine learning techniques.

Correlate 11, 14

See Acoustic correlate.

F

F₀ 3, 7, 10, 11, 12,

14 f, 16, 37, 40 f, 43

See Fundamental frequency.

Final lengthening 12, 13, 28

When the last part of a speech unit is lengthened due to grouping.

Foot 11
A rhythmic unit which builds words. The most common foot is disyllabic (consisting of two syllables) where one of the syllables is more prominent/stronger than the other syllable/s in it, which is called the stressed syllable.

Formant 11
A peak in amplitude of some frequency in a complex sound wave. A speech sound, which is a complex sound wave, has several formants and the structure of the formants is an important feature of the speech sound.

Fundamental frequency (F₀) 7
The basic frequency at which a sound vibrates. The average fundamental frequency of the adult male voice, for example, is around 120 Hz; the female voice, around 220 Hz, but F₀ is affected by for example accentuation, attitude, emphasis etc.

G

Grouping 12 f, 11, 6, 8, 9
14, 15 f, 17
A prosodic phenomenon used to group smaller units (e.g. words) into larger constituents such as prosodic phrases and utterances. Grouping also conveys the boundaries between these elements. Grouping is also called phrasing.

N

Nucleus 10
The first part of the rhyme, i.e. the vowel of the syllable. Nucleus is also called peak sometimes.

O

Onset 10, 3, 12, 27 f, 32
The initial consonant/s of the syllable.

P

PFV 20, 27-29, 31-33,
35 f, 40 f
See Prosodic feature vector.

Phone 31, 33
A token of a specific speech sound belonging to a specific phoneme. There are often several different phone variations of the phoneme, which are called allophones.

Phoneme 6, 11, 21, 29, 31 f
The segment category that the various allophones are variants of.

Pitch accent 14, 16
Accent which is conveyed with a tone/pitch.

Pitch 10, 14
A subjective, psychological sensation which is affected by the duration and intensity of the sound stimulus which has a specific frequency.

Precision 23, 25 f, 3, 33 f,
34-38, 40 f, 43

An evaluation measure that measures the precision/exactitude of a matching, e.g. classification of prominence. Viz.:

	YES is assigned	NO is assigned
YES is correct	tp	fn
NO is correct	fp	tn

$$\text{Precision} = \text{tp}/(\text{tp}+\text{fp})$$

Predictor 20 f, 16, 18, 27, 34-41

A part of the Prosodic feature vector that contains a feature's value. A prosodic feature vector contains one or more predictors per instance, e.g. word.

Predictee 20 f

A part of the Prosodic feature vector that has information about the class value, e.g. accented, for the specific instance, e.g. word. A Prosodic feature vector only contains one predictee per instance.

Prominence 11, 6-8, 9, 3, 13-17,
19-23, 25, 27 f, 31,
33, 34-41

A prosodic phenomenon used to restrain or highlight units larger than the phoneme, such as syllables, words, or phrases.

Prosodic boundary 12, 7, 16,
27, 39, 41

The boundary that separates different prosodic constituents that are results of the prosodic phenomenon of grouping.

Prosodic feature vector (PFV) 20,
16, 27

A vector containing prosodic feature values in both predictors and predictee.

Prosodic phrase break 3, 7, 13 f,
17, 20, 22,
27, 31, 33,
35, 37

See Prosodic boundary.

Prosody 9-11, 6, 3, 7 f,
13-15, 16-18, 20 f

The rhythmic, dynamic and melodic qualities of speech and is used to convey phenomena such as emphasis, intent, attitude and semantic meaning.

R

Recall 3, 23, 26, 33,
34-36, 38, 40 f, 43

An evaluation measure that measures the recall/coverage of a matching, e.g. classification of prominence. Viz.:

	YES is assigned	NO is assigned
YES is correct	tp	fn
NO is correct	fp	tn

$$\text{Recall} = \text{tp}/(\text{tp}+\text{fn})$$

Rhyme 9, 3, 27 f, 32

The last part of the syllable which consists of the nucleus/peak and coda.

S

Segment 3, 7, 13, 16,
19, 21, 29, 34

A delimited unit in the speech, e.g. phones and pauses.

Sound spectrum 11

A graph used to make acoustic analysis of a complex sound wave in which the horizontal axis represents frequency and the vertical axis the amplitude.

Speech understanding system 6, 15

A system that handles *Automatic speech understanding (ASU)*.

Speech synthesis 6, 13, 15

A technique for providing a machine with a voice which is used for producing synthetic speech. The voice can be human or synthesized.

T

Text-to-speech systems (TTS) 6, 13

A speech technology system that generates speech synthesis from texts.

Appendix 2: Program code

Program code

random.perl:

```
#!/usr/bin/perl

open(OUTFILE_SET1, ">speech_xml_jess_set_1.txt");
open(OUTFILE_SET2, ">speech_xml_jess_set_2.txt");
open(OUTFILE_SET3, ">speech_xml_jess_set_3.txt");
open(OUTFILE_SET4, ">speech_xml_jess_set_4.txt");
open(OUTFILE_SET5, ">speech_xml_jess_set_5.txt");

undef $/;

while(<>){
    @sentence_list= split(/<\s\/sentence>/);
    foreach $sentence (@sentence_list){
        $random=rand;
        if($random >= 0.8){
            print OUTFILE_SET1 $sentence;
        }
        elsif($random >= 0.6 && $random < 0.8){
            print OUTFILE_SET2 $sentence;
        }
        elsif($random >= 0.4 && $random < 0.6){
            print OUTFILE_SET3 $sentence;
        }
        elsif($random >= 0.2 && $random < 0.4){
            print OUTFILE_SET4 $sentence;
        }
        elsif($random >= 0.0 && $random <0.2){
            print OUTFILE_SET5 $sentence;
        }
    }
}

close (OUTFILE_SET1);
close (OUTFILE_SET2);
close (OUTFILE_SET3);
close (OUTFILE_SET4);
close (OUTFILE_SET5);
```

meanstand.perl:

```
#!/usr/bin/perl

open(OUTFILE2, ">ph_mean_stand_har.txt");

# Finds every row with phoneme definitions and picks the phoneme
value and msecsfors for each phoneme. 'split' scans for a " "-pattern and
splits the string into a list of substrings, returning the count of
substrings. The substrings $phoneme and $msec become a hash elements
in the last row with $msec and empty string appended to the phoneme.

while(<>){

if(/<phoneme\s\symbol=\\"(.+)\\" \s*duration=\\"(\d*)\\"\/s){

    $phoneme = $1;
    $msec = $2;
    $msecs{$phoneme} .= $msec." ";
}
}

# The output is a hash %h_msecs with sorted $phonemes. Splits the
elements in the hash and put them into an array @a_msecs. Takes
values of the $msec elements in the array and sum up the total and
counts how many phonemes there are of the same phoneme. Then it
calculates the average and prints it out.

foreach $phoneme (sort keys %msecs){
    $segments = 0;
    $total = 0;
    @msecs=split(" ", $msecs{$phoneme});
    foreach $msec (@msecs){
        $total += $msec;
        $segments++;
    }
    $average=$total/$segments;

# Standard deviation calculation

    $square_total=0;
    $square = 0;
    $stand_dev = 0;
    foreach $msec (@msecs){
        $square = ($msec-$average)*($msec-$average);
        $square_total += $square;
    }
    $stand_dev = sqrt($square_total / ($segments-1));
    printf OUTFILE2 "$phoneme\t";
    printf OUTFILE2 ("%.0f", $average);
    printf OUTFILE2 "\t";
    printf OUTFILE2 ("%.0f", $stand_dev);
    printf OUTFILE2 "\n";
}

close(INFILE2);
close(OUTFILE2);
```

prom_w_prosbreak.perl:

```
#!/usr/bin/perl

use POSIX;

open(INFILE3, "<ph_mean_stand_har.txt");

open(OUTFILE_ROUND, ">CART_INPUT/CART_input_prom_w_prosBreak_notRandom_har.txt");

# To the rhyme calc:

%all_vowels = ('A', ' ', 'E', ' ', 'I', ' ', 'O', ' ', 'U', ' ', 'Y', ' ', 'a', ' ',
              'e', ' ', 'i', ' ', 'o', ' ', 'u', ' ', 'y', ' ', 'Ä', ' ', 'Å', ' ', 'Æ', ' ',
              'Ö', ' ', 'Ø', ' ', 'ä', ' ', 'å', ' ', 'æ', ' ', 'ë', ' ', 'ö', ' ', 'ø', ' ');

%all_cons = ('C', ' ', 'D', ' ', 'G', ' ', 'J', ' ', 'L', ' ', 'N', ' ', 'S', ' ',
            'T', ' ', 'b', ' ', 'd', ' ', 'f', ' ', 'g', ' ', 'h', ' ', 'j', ' ', 'k', ' ',
            'l', ' ', 'm', ' ', 'n', ' ', 'p', ' ', 'r', ' ', 's', ' ', 't', ' ', 'v', ' ');

while(<INFILE3>){
    chomp;
    next if /^$/;
    ($sym, $dur, $dev) = split(/\t/);
    $mean_duration{$sym} = $dur;
    $standard_deviation{$sym} = $dev;
}

undef $/;

$count_word = 0;
$count_syll_total = 0;

while(<>){
    @token_list = split(/<\token>/);
    $z_score_word = 0;

    foreach $one_word (@token_list){

        next if $one_word !~ /type="word"/;
# [^\"]+ alternative to (.*)
        if ($one_word =~ /prominence="(.*?)\"/s){
            $prominence = $1;
            $one_word =~ /phrasing="(.*?)\"/;
            $break_index = $1;
            print OUTFILE_ROUND $prominence, " ";
            print OUTFILE_ROUND $break_index, " ";
            @syll_list = split(/<\syllable>/, $one_word);
            @phoneme_list_word = split(/<phon/, $one_word);
            $count_syll = 0;
            $z_score_syll = 0;
            $total_z_syll = 0;
            $z_score_last = 0;
            $z_score_ph_vow = 0;
```

```

$z_score_syll_str_one = 0;
$z_score_syll_str_two = 0;

foreach $one_syll (@syll_list){

    if($one_syll =~ /accent="\(\d+)\"/){
        $accent = $1;
        @phoneme_list = split(/<phon/, $one_syll);

        if($accent == 1 || $accent == 2){
            $count_ph_str = 0;
            $z_score_ph_str = 0;
            $total_z_ph_str = 0;

# stressed syllables' normalized values:

            foreach $one_phon_str (@phoneme_list){

                if($one_phon_str =~ /symbol="\(\.)\"/){
                    $actual_ph_str = $1;
                    ($dur_act) = $one_phon_str =~
/duration="\([0-9]+\)/";
                    $s =
$standard_deviation{$actual_ph_str};
                    $m = $mean_duration{$actual_ph_str};
                    $z_score_ph_str = ($dur_act - $m) /
$s;

                    $total_z_ph_str += $z_score_ph_str;
                    $count_ph_str ++;
                }
            }

            if($accent == 1){
                $z_score_syll_str_one = $total_z_ph_str
/ $count_ph_str;
            }

            elsif($accent == 2){
                $z_score_syll_str_two = $total_z_ph_str
/ $count_ph_str;
            }
        }
    }

#all syllables' norm values:

    $count_ph = 0;
    $z_score_ph = 0;
    $total_z_ph = 0;
    foreach $one_phon (@phoneme_list){

        if($one_phon =~ /symbol="\(\.)\"/){
            $actual_ph = $1;
            ($dur_act) = $one_phon =~
/duration="\([0-9]+\)/";

            $s = $standard_deviation{$actual_ph};
            $m = $mean_duration{$actual_ph};
            $z_score_ph = ($dur_act - $m) / $s;
            $total_z_ph += $z_score_ph;
        }
    }
}

```

```

        $count_ph ++;
    }
}
$z_score_syll = $total_z_ph / $count_ph;
$total_z_syll += $z_score_syll;
$count_syll ++;
}
}
$z_score_syll_last = 0;
$z_score_syll_rhyme = 0;
$diff_onset_rhyme = 0;
$z_score_ph_last_syll = 0;
$total_z_ph_last_syll = 0;
$count_ph_last = 0;
$total_dur_act_last_syll = 0;
$dur_act_last_syll = 0;
$dur_act_last = 0;

if($count_syll<2){
    print OUTFILE_ROUND "nil nil nil ";
}

# wordfinal syllables' norm dur:

if($count_syll > 1){

# the last index of the array @syll_list.

    $syll_last = $syll_list[$count_syll - 1];
    $z_score_ph_vow = 0;
    $z_score_ph_vow_1 = 0;
    $z_score_ph_vow_2 = 0;

    foreach $one_phon (@phoneme_list){

        if($one_phon =~ /symbol=\("(.)\)\/){
            $actual_ph_last = $1;
            ($dur_act_last_syll) = $one_phon
=~/duration=\("[0-9]+\)\/;
            $s_ph_last =
$standard_deviation{$actual_ph_last};
            $m_ph_last =
$mean_duration{$actual_ph_last};
            $z_score_ph_last_syll = ($dur_act_last_syll
- $m_ph_last) / $s_ph_last;
            $total_z_ph_last_syll +=
$z_score_ph_last_syll;
            $count_ph_last_syll ++;
        }
    }

# the rhyme and onset:

    if ($all_vowels{$actual_ph_last}){
        ($dur_act_vow) = $one_phon =~
/duration=\("[0-9]+\)\/;
        $s_vow =
$standard_deviation{$actual_ph_last};
        $m_vow = $mean_duration{$actual_ph_last};
    }
}

```

```

                                $z_score_ph_vow = ($dur_act_vow - $m_vow) /
$s_vow;

# onset duration:
# CV(C), CV(CC), CV(CCC), CV(CCCC):

                                if ($one_phon eq $phoneme_list[2]){
                                    $z_score_first_in_last = 0;
                                    $dur_act_first_in_last_cons = 0;
                                    $phon_first_in_last = $phoneme_list[1];
                                    ($actual_ph_first_in_last) =
$phon_first_in_last =~ /symbol="\(.)\"/;
                                    $z_score_onset_in_last
=&z_score_onset($phon_first_in_last);
                                    sub z_score_onset{

                                        if($all_cons{$actual_ph_first_in_last}){
                                            my $phon_first_in_last =
shift(@_);
                                            ($dur_act_first_in_last_cons) =
$phon_first_in_last =~ /duration="\([0-9]+\)/;
                                            $s_last =
$standard_deviation{$actual_ph_first_in_last};
                                            $m_last =
$mean_duration{$actual_ph_first_in_last};
                                            $z_score_onset_in_last
=(($dur_act_first_in_last_cons - $m_last) / $s_last);
                                        }
                                    }
                                }

# CCV(C), CCV(CC), CCV(CCC), CCV(CCCC)

                                elsif($one_phon eq $phoneme_list[3]){
                                    $z_score_first_in_last = 0;
                                    $dur_act_first_in_last_cons = 0;
                                    $phon_sec_in_last = $phoneme_list[2];
                                    $phon_first_in_last = $phoneme_list[1];
                                    ($actual_ph_sec_in_last) =
$phon_sec_in_last =~ /symbol="\(.)\"/;
                                    ($actual_ph_first_in_last) =
$phon_first_in_last =~ /symbol="\(.)\"/;
                                    $z_score_first_in_last =
&z_score_onset($phon_first_in_last);
                                    $z_score_sec_in_last =
&z_score_onset($phon_sec_in_last);
                                    $z_score_onset_in_last =
$z_score_sec_in_last + $z_score_first_in_last;
                                }

# CCCV(C), CCCV(CC), CCCV(CCC), CCCV(CCCC)

                                elsif($one_phon eq $phoneme_list[4]){
                                    $z_score_first_in_last = 0;
                                    $dur_act_first_in_last_cons = 0;
                                    $phon_first_in_last = $phoneme_list[1];
                                    $phon_sec_in_last = $phoneme_list[2];

```

```

        $phon_third_in_last = $phoneme_list[3];
        ($actual_ph_third_in_last) =
$phon_third_in_last =~ /symbol="\(.)\"/;
        ($actual_ph_sec_in_last) =
$phon_sec_in_last =~ /symbol="\(.)\"/;
        ($actual_ph_first_in_last) =
$phon_first_in_last =~ /symbol="\(.)\"/;
        $z_score_first_in_last =
&z_score_onset($phon_first_in_last);
        $z_score_sec_in_last =
&z_score_onset($phon_sec_in_last);
        $z_score_third_in_last =
&z_score_onset($phon_third_in_last);
        $z_score_onset_in_last =
$z_score_third_in_last + $z_score_sec_in_last +
$z_score_first_in_last;
    }
}

# last index of the array @phoneme_list:

$phon_last = $phoneme_list[$#phoneme_list];
$z_score_last_1 = 0;
$z_score_last_2 = 0;
$dur_act_last_cons = 0;
($actual_ph_last) = $phon_last =~ /symbol="\(.)\"/;

if($all_cons{$actual_ph_last}){
    ($dur_act_last_cons) = $phon_last =~
/duration="\([0-9]+\)/;
    $s_last = $standard_deviation{$actual_ph_last};
    $m_last = $mean_duration{$actual_ph_last};
    $z_score_last = ($dur_act_last_cons - $m_last) /
$s_last;
}
$z_score_syll_last = $total_z_ph_last_syll /
$count_ph_last_syll;
print OUTFILE_ROUND $z_score_syll_last;
print OUTFILE_ROUND " ";

if($z_score_last != 0){
    $z_score_syll_rhyme = ($z_score_last +
$z_score_ph_vow) / 2;
    print OUTFILE_ROUND $z_score_syll_rhyme, " ";
    print OUTFILE_ROUND " ";
}

elsif($z_score_last == 0){
    $z_score_syll_rhyme = $z_score_ph_vow;
    print OUTFILE_ROUND $z_score_syll_rhyme, " ";
    print OUTFILE_ROUND " ";
}
$diff_onset_rhyme = ($z_score_onset_in_last -
$z_score_syll_rhyme);
print OUTFILE_ROUND $diff_onset_rhyme;
print OUTFILE_ROUND " ";
}

```

```

$count_word ++;

if($z_score_syll_str_one == 0){
    print OUTFILE_ROUND $z_score_syll_str_two;
    print OUTFILE_ROUND " ";
}

elseif($z_score_syll_str_one != 0){
    print OUTFILE_ROUND $z_score_syll_str_one;
    print OUTFILE_ROUND " ";
}
$count_syll_total += $count_syll;
if (@token_list[$count_word-1] =~ /type="pause"/ &&
@token_list[$count_word] =~ /type="word"/ ){
    print OUTFILE_ROUND "PAUSE_B ";
}
else{
    print OUTFILE_ROUND "NO_PAUSE_B ";
}
if (@token_list[$count_word+1] =~ /type="pause"/ &&
@token_list[$count_word] =~ /type="word"/ ){
    print OUTFILE_ROUND "PAUSE_A ";
}
else{
    print OUTFILE_ROUND "NO_PAUSE_A ";
}
$total_z_ph_word = 0;
$z_score_ph_word = 0;
$count_ph_word = 0;

foreach $one_phon_of_word (@phoneme_list_word){

    if($one_phon_of_word =~ /symbol="(.)\"/){
        $actual_ph_word = $1;
        ($dur_act) = $one_phon_of_word =~
/duration="([0-9]+)\"/;
        $s = $standard_deviation{$actual_ph_word};
        $m = $mean_duration{$actual_ph_word};
        $z_score_ph_word = ($dur_act-$m) / $s;
        $total_z_ph_word += $z_score_ph_word;
        $count_ph_word ++;
    }
}
$z_score_word = $total_z_ph_word / $count_ph_word;
print OUTFILE_ROUND $z_score_word;
print OUTFILE_ROUND "\n";
}
}

print $count_syll_total, "\n\n";
print $count_word, "\n\n";

close(INFILE3);
close(OUTFILE_ROUND);

```

annotate_prom.perl

```
#!/usr/bin/perl -w

open(INFILE_1,
"<CART_OUTPUT/CART_output_prom_w_prosBreak_notRandom_har_class.txt")
;

while(<INFILE_1>)
{
    next if /^$/;
    chomp;
    ($pred_label) = $_ =~ / (.+)\$/;
    push (@prom_label, $pred_label);
}

while(<>){

    if ( /type=\"word\"/ ){
        $isWord = 'yes';
    }
    if ( /type=\"(delimiter|pause)\"/ ){
        $isWord = 'no';
    }

    if ( /prominence/ && $isWord eq 'yes' ){
        $label = shift @prom_label;
        s/prominence=\"(.*)\"/prominence=\"$1:$label\"/;
    }
    print;
}

close(INFILE_1);
```

eval_prom.perl:

```
#!/usr/bin/perl -w

$count_deacc_deacc=0;
$count_acc_acc=0;
$count_prom_prom=0;
$count_prom_deacc=0;
$count_acc_deacc=0;
$count_prom_acc=0;
$count_acc_prom=0;
$count_deacc_prom=0;
$count_deacc_acc=0;

while(<>){

    if ( /type=\"word\"/ ){
        $isWord = 'yes';
    }
    if ( /type=\"(delimiter|pause)\"/ ){
        $isWord = 'no';
    }

    if ( /prominence=\"deacc:deacc/ && ( $isWord eq 'yes' ) ){
        $count_deacc_deacc ++;
    }
    elsif(/prominence=\"deacc:acc/ && ( $isWord eq 'yes' ) ){
        $count_deacc_acc ++;
    }
    elsif(/prominence=\"deacc:prom/ && ( $isWord eq 'yes' ) ){
        $count_deacc_prom ++;
    }
    elsif(/prominence=\"acc:acc/ && ( $isWord eq 'yes' ) ){
        $count_acc_acc ++;
    }
    elsif(/prominence=\"acc:deacc/ && ( $isWord eq 'yes' ) ){
        $count_acc_deacc ++;
    }
    elsif(/prominence=\"acc:prom/ && ( $isWord eq 'yes' ) ){
        $count_acc_prom ++;
    }
    elsif(/prominence=\"prom:deacc/ && ( $isWord eq 'yes' ) ){
        $count_prom_deacc ++;
    }
    elsif(/prominence=\"prom:acc/ && ( $isWord eq 'yes' ) ){
        $count_prom_acc ++;
    }
    elsif(/prominence=\"prom:prom/ && ( $isWord eq 'yes' ) ){
        $count_prom_prom ++;
    }
}

print "      deacc  acc  prom\ndeacc ";
print $count_deacc_deacc, "      ";
print $count_deacc_acc, "      ";
print $count_deacc_prom, "\nacc      ";
print $count_acc_deacc, "      ";
print $count_acc_acc, "      ";
```

```
print $count_acc_prom, "\nprom ";
print $count_prom_deacc, " ";
print $count_prom_acc, " ";
print $count_prom_prom, "\n";
```