Barbertje Streefkerk

Prominence
Acoustic and lexical/syntactic correlates

ACLC

: LOT
Netherlands
Graduate
School of
Linguistics

Landelijke Onderzoekschool Taalwetenschap

PROMINENCE

ACOUSTIC AND LEXICAL/SYNTACTIC CORRELATES

Barbertje Marieke Streefkerk

Published by
LOT
Trans 10
3512 JK Utrecht
The Netherlands

phone: +31 30 253 6006
fax: +31 30 253 6000
e-mail: lot@let.uu.nl
http://wwwlot.let.uu.nl/

Cover illustration: door of a bus in Amsterdam

ISBN 90-76864-19-5
NUGI 941

PROMINENCE

ACOUSTIC AND LEXICAL/SYNTACTIC CORRELATES

ACADEMISCH PROEFSCHRIFT

ter verkrijging van de graad van doctor
aan de Universiteit van Amsterdam
op gezag van de Rector Magnificus
prof. mr. P.F. van der Heijden
ten overstaan van een door het college voor promoties ingestelde commissie, in het
openbaar te verdedigen in de Aula der Universiteit
op dinsdag 8 oktober 2002, te 14.00 uur
door

Barbertje Marieke Streefkerk

geboren te Amsterdam

Promotor: prof. dr. L.C.W. Pols (Universiteit van Amsterdam)

Co-promotor: dr. L.F.M. ten Bosch (Katholieke Universiteit Nijmegen)

Overige leden: dr. A. Batliner (Universität Erlangen-Nürnberg)
prof. dr. L. Boves (Katholieke Universiteit Nijmegen)
prof. dr. V.J.J.P. van Heuven (Universiteit Leiden)
prof. dr. J-P. Martens (Universiteit Gent)
prof. dr. R.J.H. Scha (Universiteit van Amsterdam)

Faculteit der Geesteswetenschappen

voor mijn ouders

DANKWOORD

Bij het tot stand komen van dit proefschrift hebben veel mensen op heel veel verschillende manieren een bijdrage geleverd.

Ik wil al deze mensen mijn dank uitspreken.

Mijn begeleiders Louis ten Bosch en Louis Pols voor de uitvoerige discussies op vrijdagmiddag, die ik als positief heb ervaren, en die altijd weer nieuwe ideeën naar voren brachten.

De commissieleden Anton Batliner, Lou Boves, Vincent van Heuven, en Jean-Pierre Martens voor hun uitvoerige en nuttige commentaar en hulp om dit toch nog in een laatste fase te verwerken. Ook wil ik hier Johan Matter hartelijk danken voor zijn steun en hulp bij het afmaken van mijn proefschrift en voor de gezellige middagen met de asbak in het midden.

Mijn collega's van het Instituut Corina van As, Ellen Berkman, Paul Boersma, Jan van Dijk, Ineke van den Dikkenberg-Pot, Karijn Helsloot, Dirk Jaasma, Florien Koopmans-van Beinum, Rob van Son, Jeannette van der Stelt en Ton Wempe waarmee ik vijf jaar met veel plezier heb samengewerkt en elke ochtend om kwart over tien koffie heb gedronken.

Mijn promovendi-collega's Annerieke, Ceske, Hedde, Jasper, Margot, Nel, Victoria, en Wim voor de wekelijkse lunch en de strijd tegen het bursalenbestaan van promovendi aan de UvA.

Frans en Willy, mijn ouders, Kees en Maartje, mijn broer en zus die altijd achter mij stonden.

En mijn paranimfjes, mijn huisgenoten, en vriendinnen Freke en Caroline voor alles wat jullie voor mij gedaan hebben.

Carla, Lilly, Simone en Yvonne voor hun motiverende gespreken, voor hun morele ondersteuning en het lekkere eten.

Vooral Christian, mijn lief, zonder wie dit proefschrift niet tot stand zou zijn gekomen.

Barbertje Streefkerk

TABLE OF CONTENT

1 Introduction 1
1.1 Notion of prominence 2
1.1.1 General viewpoint 2
1.1.2 Phonetic viewpoint 2
1.1.3 Linguistic viewpoint 3
1.2 Topic of investigation 6
1.3 Usefulness of acoustic and lexical / syntactic correlates of prominence 8
1.4 Research method 9
1.5 Outline of this study 9

2 Prominence marking by naive listeners 11
2.1 Introduction 12
2.1.1 Literature survey 12
2.1.2 Our approach 13
2.2 Speech material 15
2.2.1 Recordings 16
2.2.2 Material 16
2.2.3 Speakers 16
2.2.4 Selected speech material 17
2.3 Two pilot experiments to define prominence 17
2.3.1 Method 17
2.3.2 Results 18
2.3.3 Conclusion 23
2.4 Main experiment on assigning prominence 23
2.4 1 First set marked by ten listeners 23
2.4.1.1 Resulting prominence marks 24
2.4.1.2 Differences within and between listeners 25
2.4.1.3 Clustering of the cumulative prominence marks 29
2.4.2 Second set marked by one 'optimal' listener 33
2.4.2.1 Prominence assignment 33
2.5 Concluding remarks 34

3 Lexical and syntactic correlates of prominence 37
3.1 Introduction 38
3.1.1 Relevant studies 40

3.1.1.1 Data driven studies 40
3.1.1.2 Application-oriented studies 41
3.2 Pilot study to find lexical / syntactic correlates 44
3.3 Main experiment on lexical / syntactic correlates of prominence 45
3.3.1 Assigning lexical and syntactic features 46
3.3.1.1 Description and evaluation of the automatically derived Part-of-Speech (POS) tags 48
3.3.2 Relationship between word class and prominence 49
3.3.3 Relationship between word length and prominence 53
3.3.4 Relationship between the position of a word in a sentence and prominence 56
3.3.5 Adjective-Noun combinations and prominence 56
3.3.6 Algorithm to predict prominence on lexical / syntactic input for prominence prediction 58
3.4 Independent test of the prominence assignment rules 67
3.5 Discussion and conclusion 68

4 Acoustical correlates of prominence 71
4.1 Introduction 72
4.1.1 General description of possible acoustic correlates 72
4.1.2 Relevant studies on automatic feature extraction 74
4.2 Feature extraction 77
4.2.1 Segmentation and labeling of the speech material 77
4.2.1.1 Training of the HMM-recognizer 78
4.2.1.2 Resulting segmentation 79
4.2.1.3 Accuracy of the automatic segmentation 80
4.2.2 Acoustical correlates 81
4.2.2.1 Unit selection 81
4.2.2.2 F_0 features 82
4.2.2.2.1 Measuring F_0 84
4.2.2.2.2 Extracting F_0 features 86
4.2.2.3 Duration features 89
4.2.2.4 Intensity features 96
4.2.2.4.1 Measuring intensity 96
4.3 Summary and conclusion 100

5 Neural net classification of prominence with acoustic input features 101
5.1 Introduction 102
5.1.1 How feed-forward networks work 102
5.1.1.1 General training procedure 103
5.1.2 Distributions of prominence classes and the relationship to applications 105
5.2 Prominence recognition with neural networks 105
5.2.1 Acoustical input features 105
5.2.2 Pre-processing of the input features 107

5.2.2.1 Correlation 107
5.2.3 Design of the training and testing data 108
5.2.4 Binary prominence classification 110
5.2.4.1 Testing with the Independent test set 113
5.2.4.2 Summary and conclusion 114
5.2.5 Gradient prominence prediction 115
5.2.5.1 Results 116
5.2.6 Analyses of individual features 119
5.2.6.1 Analyzing the neural network 119
5.2.6.2 Analyzing the performance of the individual feature 123
5.2.6.3 Summary and conclusion 125
5.2.7 Analyzing combinations of features 125
5.2.8 Prominence classification with an 'optimal' feature combination 126
5.3 Discussion and conclusion 127

6 General conclusion and discussion 129
6.1 Introduction 130
6.2 Prominence assignment by naive listeners 130
6.2.1 Word or syllable prominence marking 130
6.2.2 Binary or gradient prominence marking 131
6.2.3 Consistency and reliability 132
6.2.4 Concluding remarks 132
6.3 Lexical / syntactic correlates of prominence 132
6.3.1 Individual features 133
6.3.2 Prominence prediction on textual input 133
6.3.3 Discussion about lexical / syntactic correlates 134
6.3.3.1 Comparison to the literature 134
6.3.3.2 Method used 134
6.3.3.3 Useful for Text-to-Speech 135
6.4 Acoustical correlates of prominence 136
6.4.1 Individual features 136
6.4.2 Prominence prediction on acoustic input 136
6.4.2.1 Complexity 137
6.4.3 Discussion about acoustic correlates 137
6.4.3.1 Comparison to the literature 137
6.4.3.2 Method used 138
6.4.3.3 HMM-alignment 138
6.4.3.4 Normalizations 138
6.4.3.5 Separate use of linguistic and acoustic features 139
6.5 Future research 139
6.5.1 Promising features of prominence 139
6.5.2 Speech-technology applications 139
6.5.3 Combining linguistic and acoustical features 140

Appendix A 2.1 Correspondence matrix assigning prominence without pitch movements 143
Appendix A 2.2 NIST header 144
Appendix A 2.3 Demographic data of the training set 145
Appendix A 2.4 Demographic data of the test set 146
Appendix A 2.5 Instruction for the listening experiment 147
Appendix A 3.1 Overview of predicted prominence marks and lexical and syntactic correlates 148
Appendix A 4.1 SAMPA symbols 149
Appendix A 5.1 Scaling values 150

Bibliography **151**

Summary **161**

Samenvatting **165**

Curriculum vitae **169**

1

INTRODUCTION

Abstract

This study investigates the acoustic and linguistic correlates of prominence. In this chapter the notion of prominence is explained and its use in language and communication is illustrated. Next, the research questions that will be dealt with in this study, will be identified.

1.1 Notion of prominence

This chapter is divided into five sections. In the first section, we will explain what we consider 'prominence' to be from a general viewpoint, as well as from a phonetic and a linguistic viewpoint. The things we want to know about prominence are dealt with in the second section. In the third section we will motivate why we want to investigate prominence and in the fourth section we will discuss how prominence will be investigated. The final section will present an outline of the present study.

1.1.1 General viewpoint

When we listen to speech some parts seem more prominent than others. In other words, we perceive specific parts of the speech signal as uttered with more 'emphasis' than other parts. This emphasis is called 'prominence'.
Prominence is not a fixed property. It changes over time and is dependent on many linguistic, textual and acoustic-phonetic factors. Word groups, single words, syllables and even single phonemes can differ in prominence (Ladd, 1996; Sluijter, 1995; van Heuven, 1994). This difference in prominence is not a binary property but rather a gradient property (Terken, 1996; Rietveld & Gussenhoven, 1985).
In many languages, such as Dutch, English, German and French prominence is used primarily to structure a message i.e. to give emphasis to specific parts of the message. Prominence is just one of the ways in which the information structure of a message can be made more explicit; another way is phrasing. One could also change the word order such as with clitic pronouns in French (je, tu, il) that can never receive prominence and therefore give rise to structures such as c'est moi qui l'ai fait 'it was I who did it' (cf. *Je l'ai fait, I did it).
Structuring the message is not the only benefit of prominence; applying appropriately varying levels of prominence also increases the naturalness and the comprehensibility of speech.
A speaker uses prominence to mark those parts that are important in his message, and the listener uses (perceived) prominence in order to know which parts are of special interest for the perceived message. The listener combines bottom-up information from the speech signal with his expectation of prominence on the basis of his knowledge of the language (top-down information). In this study, we will concentrate on the *perceived* prominence.

1.1.2 Phonetic viewpoint

From a phonetic viewpoint the notion of prominence is not clearly defined. Pitch accent, sentence accent, stress, lexical (word) stress, word stress, reduction of vowels or syllables are all terms for which the definition may vary between linguistic models, but all are related to prominence. These terms often describe

nearly the same phenomena and therefore may lead to confusion. In this study we restrict ourselves to use two terms: pitch accent and lexical (word) stress.

Pitch is strongly related to F_0, referring to the periodicity in a harmonic complex. An increase in F_0 correlates with an auditory sensation of a higher pitch. Changes in F_0 closely correspond with perceived pitch movements. The intonation contour can be seen as consecutive pitch movements of which some can be associated with pitch accents. These phenomena are described on a more abstract level by using so-called intonation grammars of which two important ones are the IPO intonation grammar and the auto-segmental approach (TOBI). The IPO grammar, deals with subsequent rising and falling pitch movements ('t Hart et al., 1990); the auto-segmental grammar (Silverman et al., 1992; Gussenhoven, 1984; Pierrehumbert, 1980) describes pitch movements in a functional and abstract way. (See for more information section 2.1.1.)

In many languages lexical (word) stress is a property of a syllable within the domain of a word and is generally defined in the lexicon for languages such as Dutch, German and English. Lexical (word) stress can be seen as a linguistic phenomenon (this will be described in the next subsection). However, if it concerns the acoustic realization of a word or its percept it is evidently more related to phonetic-acoustic properties of the speech signal.
The usual distinction in lexical (word) stress is between 'stressed' and 'unstressed' syllables, but a distinction of four degrees (1) 'primary', (2) 'secondary' (3) 'tertiary' and (4) 'weak' is also used. In acoustic phonetics lexical (word) stress is in many languages usually due to an increase in intensity of the stressed syllable, but increase in duration and F_0 changes may be involved as well (Lehiste, 1970; Fry 1958).

In this study we do not concentrate so much on the actual relationship between prominence on the one hand and lexical (word) stress and pitch accent on the other. We merely want to say that these phenomena are closely related to prominence and that prominence is a complex mix of several phenomena.

1.1.3 Linguistic viewpoint

From a linguistic viewpoint prominence is mainly concerned with focus and lexical (word) stress as a property of the lexicon. From a linguistic viewpoint the function of lexical word stress is probably a more indirect way to compartmentalize the mental lexicon of the listener (Cutler, 1984). If the listener knows the position of the lexically stressed syllable in words this may help to recognize the word more quickly in the appropriate sublexicon. The actual realization of lexical (word) stress in the acoustic signal is related to the prominence of the syllable. By placing

prominence on different syllables, the phoneme string can acquire widely differing meanings.
This is shown in the following examples (prominent syllables are indicated by capital letters):

i. *CAmeraatje (little camera)*
ii. *kameRAAdje (little companion)*
iii. *CAnon (canon)*
iv. *KaNON (cannon)*

Only the perceived differences of prominence of the first versus the second or third syllable disambiguate between the different meanings of the two words in the examples above. The first syllable is perceived as more prominent in (i) and (iii) and the third / second syllable is perceived as more prominent in (ii) and (iv). This is related to lexical (word) stress.

We consider an utterances to have an 'information structure' that is related to the relative prominence of all the speech elements. Analyzing this information structure is complex and sometimes controversial, however, it is closely related to prominence.
In as far as information structure is concerned it is common to distinguish between 'given' and 'new' information (e.g. Halliday, 1967). 'Given' refers to information already supplied by the previous linguistic context whereas 'new' information has not been previously supplied.
In this context, some authors speak of 'focus'. The speaker can highlight the information for the listener that is at the 'focus' of their communicative interest (Baart, 1987; Nooteboom & Kruyt, 1987; Gussenhoven, 1984; Ladd, 1980).

We will describe the following examples in terms of focus. Distinctions such as 'broad' and 'narrow focus' or 'contrastive focus' can be made. We decided to explain in our example sentences only '(narrow) focus' (A) and 'contrastive focus' (B). The last example shows a mix of lexical and focal contracts (C), resulting in completely different meanings of the sentences.

A)
i. *IK wil nog twee bloemen*
ii. *Ik wil NOG twee bloemen*
iii. *Ik wil nog TWEE bloemen*
iv. *Ik wil nog twee BLOEMen*

Changes in prominence patterns can guide the attention of the listener to specific words. Different words are prominent in sentences (i) to (iv). A neutral translation is: *I want two more flowers.* If the word *ik (I)* is more prominent, the attention is

guided to this word and for the listener it is clear that it is 'me' and not someone else who wants the flowers. In (ii) and (iii) there is a difference of meaning in the sentence; (ii) means *I want two additional flowers* and (iii) means *I want exactly two more flowers,* rather than four more. In (iv) the sentence transmits the information that the speaker wants flowers rather than something else.

When contrast is required, the speaker can highlight different parts of a sentence, which makes the contrast more recognizable.

B)

v. *Ik ga niet naar ZAANdam maar naar LEERdam (I do not go to Zaandam but to Leerdam)*

vi. *Het is niet DE boek maar HET boek (it is not the (non-neuter) book, but the (neuter) book)*

In (B-v) the first syllable of the names of two Dutch cities are put in contrast. Normally the lexical stress of these two names is located on the last syllable, so this is an example were two normally not lexically stressed syllables are more prominent than the lexically stressed one. The last example (vi) shows that the Dutch articles *de* and *het* can also be put into contrast and can be more prominent than the other words in the phrase.

C)

Just as word meaning can change with alterations to syllable prominence, so can the meaning of a sentence alter as prominence is given to differing words.

i. *uitsluitend VOOR instappen (only get on at front)*

ii. *uitsluitend voor INstappen (only for getting on)*

iii. *naTUURlijk(,) VOORkomen van bosbrand is wenselijk (of course the occurrence of forest fire is desirable)*

iv. *naTUURlijk(,) voorKOMen van bosbrand is wenselijk (of course the prevention of forest fire is desirable)*

v. *Natuurlijk VOORkomen van bosbrand is wenselijk (the natural occurrence of forest fire is desirable)*

vi. *Natuurlijk voorKOMen van bosbrand is wenselijk (the natural prevention of forest fire is desirable)*

Examples (C-i) and (C-ii) are written on each door of the Amsterdam busses and streetcars except on the first carriage. The meaning of the whole sentence depends on differences in the allocation of prominence. If the word *voor* is more prominent than the word *instappen*, as in (i), incoming passengers should enter the streetcar

only at the front door. In the second reading (ii) *instappen* is more prominent; the meaning then is that this door should only be used to enter and not to exit.

Four different meanings are possible in examples (C iii-vi), because of lexical ambiguity as well as sentence ambiguity. In (iii) and (iv) there is a lexical conflict of different lexical meanings of the word *voorkomen*. Prominence on the first syllable means *occurrence*, but with prominence on the second syllable it means *prevention*. In (iii) and (iv) versus (v) and (vi) the difference in prominence of the first and the second word disambiguates the meaning of the sentence. *Natuurlijk* in (iii) and (iv) means *of course* and in (v) and (vi) *natural*. Thus four different meanings are possible, (iii) means *of course the occurrence of forest fire is desirable* and (iv) means *of course prevention of forest fire is desirable*. A different phrasal meaning is given, for (v) and (vi), (v) means *natural occurrence of forest fire is desirable* and a prominence pattern as in (vi) changes the meaning to *natural prevention of forest fire is desirable*.

The main topic in this study is not to describe the relationship between prominence and either focus, or given and new information (see for that for instance van Donzel, 1995). It also does not study prominence in terms of pitch accent, using the IPO intonation grammar ('t Hart et al., 1990), or the auto-segmental theory (Gussenhoven, 1984; Pierrehumbert, 1980). We concentrate on prominence as such. These and other topics will be discussed and compared with the literature in the introductions to the appropriate chapters. Prominence gives access to the information structure and is closely related to concepts of pitch accent and lexical (word) stress.

1.2 Topic of investigation

The main topic in this study will be prominence itself. Prominence refers to the degree in which a phoneme, syllable, and / or a word is perceived to stand out from its environment. Prominence is therefore primarily a perceptual concept. On the one hand the listener uses variations in length, loudness and pitch as cues (bottom-up information) in order to signal relative prominence of a unit (Terken, 1996; Hermes & Rump, 1994; Rietveld, 1983; Lehiste, 1970). On the other hand the listener's perception is biased by expectations. These expectations are based on knowledge of the language (top-down information). Examples of linguistic knowledge are the syntax of a language, the differences between content word and function word, Part-of-Speech information in general, position of a word in a sentence, and word frequency (Altenberg, 1987; Baart, 1987; Chomsky & Halle, 1968). The relationships are visualized in figure 1.1.

Our research questions fold into three parts:

1) How to find an operational definition of prominence?
2) What are the linguistic determinants / correlates of prominence (top-down)?
3) Which acoustic correlates contributes to the perception of prominence (bottom-up)?

The first question concerns the perceptual notion of prominence and how it should be defined. An operational definition is needed in order to label a database in terms of prominence. General questions arise here. Is it necessary to have experts listeners label prominence, or can naive native listeners do this labeling as well? On which unit (segment, syllable, word) should the judgments be given and should one use one or more subjects for this labeling task? Prominence is a relative and gradient phenomenon; should labeling thus be multi-valued or binary? If multi-valued what should be the range of the scale; a 10-point scale or a 4-point scale? Once labeled, what are the consistency and the reliability of the labelers? All these questions are discussed in chapter 2.
The second research question deals with the linguistic determinants / correlates of prominence. Lexical and syntactic features such as Part-of-Speech, word length and

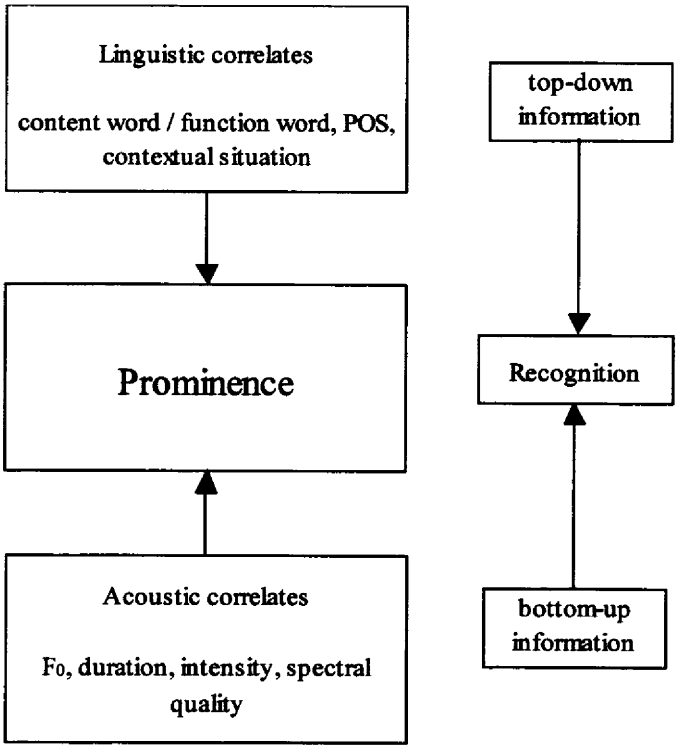


Figure 1.1: Relation of prominence and linguistic and acoustic correlates.

position of a word in the sentence are related to pitch accents (Hirschberg, 1993; Baart, 1987) and therefore to perceived prominence. Do these features correlate with perceived prominence and if so, how? To what extent do these features contribute to the prediction of prominence purely on such linguistic information? Chapter 3 provides more information on this topic.

The third research topic concerns the acoustic correlates of perceived prominence. What are the acoustic correlates of prominence? From literature it is known that F_0 changes, and duration of vowels and / or syllables are related to lexical (word) stress and pitch accent. What is the relationship of these acoustic correlates to prominence, and to what extent are combinations of these features correlates of prominence? Can the prominence labels from listeners be 'predicted' by a classifier that is only using the acoustic signal as input? What is the contribution of a selected set of acoustic features to prominence classification? And to what extent do certain normalizations, for instance intrinsic vowel duration and speaking rate, contribute to a better prominence prediction. Chapter 4 and chapter 5 discusses these questions.

1.3 Usefulness of acoustic and lexical / syntactic correlates of prominence

From a general viewpoint the perceptual phenomenon prominence seems to have an important communicative function. Therefore it is interesting from a phonetic viewpoint to investigate prominence itself as a perceptual phenomenon and correlates of prominence. The following questions form an interesting topic. To what extent is prominence marking by naive listeners useful for our research? What is the consistency and reliability of listeners? Which acoustic correlates and linguistic determinants contribute to the perception of prominence and what is their effect if they are used to predict prominence solely from acoustic input features on the one hand and from linguistic correlates on the other?
Apart from dealing with the communication process between speaker and listener this study is also concerned with speech technology. This introduces specific limitations to the investigation method. For example, the acoustic features on which the automatic classification of prominence will be based must be derived from the speech signal in an automatic way without additional human (knowledge-based) correction or intervention.

Three applications for speech technology are briefly introduced here.
The first application concerns prominence prediction for a Text-to-Speech system to increase the naturalness and intelligibility of synthetic speech. Most speech synthesis systems today use the notion of pitch accent and lexical (word) stress without using the degree of prominence. Predicting prominence for speech synthesis purposes has to be solely based on textual input. The prediction of prominence is not a unambiguous process and the location and degree of prominence is not explainable from textual information only. Reminiscent to Bolinger's remark about humans not

being a 'mind-reader' (Bolinger, 1972), a computer is certainly not. This makes that most of pragmatic and semantic information can not play a role and will not be used in the experiments of the present study.

The second and third application are in the field of automatic speech recognition: a prominence indicator and an algorithm for disambiguating the meaning of sentences. Prominence can guide or alter the meaning of the sentence, as, for instance, illustrated in the example given before *Ik wil nog twee bloemen.* A prominence indicator or classifier can provide useful information for the speech recognition process, more specifically during the word search process. Knowing the degree of prominence can help to decide whether a word is important for communication. Prominence indication and sentence disambiguation can be based on acoustic input, as well as on information coded in the lexicon.

1.4 Research method

The approach we choose to investigate the perceptual notion of prominence and the related acoustic and linguistic correlates, imposes restriction on our research methods as well as on the choice of speech material to be investigated.
First, the speech material should constitute a sufficiently large speech corpus that is valid for speech technological applications.
Second, the prominence labeling should still be possible for such a large corpus, while leaving us with as much detailed labeling as possible. With respect to this constraint, the use of naive listeners is of special interest.
Third, feature extraction and prediction should be done automatically. However, we still want to control the feature extraction and want to describe and to analyze the features in an interpretable way. The prediction of prominence should also be controllable. Once knowing the 'optimal' features, rules should be formulated and used for prominence prediction, either on acoustic or linguistic input. The analysis of the linguistic and acoustic features should not be based on a purely statistical and / or brute force approach, because we want to control the individual features and their contribution to the prediction of prominence. This limits the prediction tools to simple ones. The prominence prediction should be tested on an independent test set in order to get an idea of the consistency of the prediction and the generalization capability of the prominence classifier.

The appropriate literature will be discussed in detail in the separate chapters.

1.5 Outline of this study

This study deals with the question which features in the speech signal and in the text (acoustic, linguistic) can be used to predict prominence automatically. To this end a

test and a training corpus were designed for classification. A group of listeners judged the speech material used and marked all the sentences for prominence (chapter 2). In chapter 3 the lexical / syntactic correlates are analyzed and are used to predict prominence by using textual features. This mainly concerns correlates / determinants such as word class, and the position of words in the sentence. Chapter 4 focuses on the acoustic correlates that are used to predict prominence. In contrast to chapter 3, the emphasis in chapter 4 is on the acoustic features that can be automatically derived from the speech signal. In addition, this study will attempt to discover which acoustic features can be used for the automatic classification and prediction of prominence (chapter 5). A general discussion as well as conclusions and ideas for further research will be presented in the last chapter.

2

PROMINENCE MARKING BY NAIVE LISTENERS [1]

Abstract

After a short literature survey, an operational definition of prominence is developed on the basis of two pilot experiments (81 sentences). The Dutch Polyphone speech material used is described. Ten listeners marked word prominence for the training material (1244 sentences) to be used later. A detailed discussion of the behavior of the listeners and the consistency and reliability of listeners is presented. Finally, the design of the independent test material (1000 sentences) is described.

[1] Parts of this chapter were published in Streefkerk et al. (1997) and Streefkerk & Pols (1998).

2.1 Introduction

This chapter deals with the marking of 'prominent' parts in speech utterances. Prominence marking looks easy, but it is rather complex. The notion of prominence is not clearly defined and various ways to mark prominence are used in the literature. In the next subsection various approaches are discussed.

2.1.1 Literature survey

In the literature, the notion of prominence is made operational in various ways. Mainly we deal with the following options: 1) mark every syllable, 2) mark the prominent syllable, 3) mark every word, 4) mark the prominent word, 4) mark binary, 6) or on a gradient scale. Portele & Heuft (1997) and Fant & Kruckenberg (1989) used a 31-point scale: each syllable is judged for the amount of prominence by, in principle, one listener. In this method the judgments are very finely tuned, and the listener must listen to every syllable in detail. This method is only possible for small samples of speech material.
Grover et al. (1997) used initially a 'unlimited' scale for word prominence, which was immediately reduced to a 10-point scale. In Grover et al. (1998), in which a prosodic database is described, this reduced scale is maintained. Listeners were asked to rank all individual words in the utterance for prominence. Marking the words is more efficient for large databases than marking every syllable for prominence, but still every single word has to be marked. Strangert & Heldner (1995) mark word prominence on a 4-point scale.
In the research of Kießling (1996) prominence was binary marked by ten listeners. The listeners were asked to mark only the prominent words, so not for all words a judgment was given. Kießling (1996) used the cumulative prominence marks of the listeners as an indication of the prominence degree. This way of marking word prominence is efficient for labeling large databases.

In a more indirect approach, the prominence differences related to various pitch accents are studied with the help of speech resynthesis (Gussenhoven et al., 1997; Rump & Hermes, 1996; Terken, 1996; Terken, 1991; Gussenhoven, 1985; Rietveld & Gussenhoven, 1985). This indirect approach concentrates on the notion of pitch accent and its perceptual variation of prominence. Terken (1996) vary the peak heights of pitch accents and present the manipulated speech to listeners who had to judge, mostly on a 10-point scale, the degree of prominence of the syllables or the words. One of the findings is that words at the beginning of an utterance must have a larger peak height than words at the end of an utterance, in order to be perceived with the same prominence.
The notion of pitch accent is more or less reduced in these studies to changes in F_0. One has tried to devise grammars to describe these phenomena. The IPO grammar of intonation e.g. deals with subsequent rising and falling pitch movements ('t Hart et al., 1990), whereas the auto-segmental intonation grammar (Silverman et al.,

1992; Gussenhoven, 1984; Pierrehumbert, 1980) describes pitch movements in a functional way.
In the IPO intonation grammar, pitch movements are defined as being accent lending and the others mark boundaries. A typical pitch movement is defined by the shape of the movement and by the onset in the syllable. Accent lending is interpreted in this study as closely related to prominence.
A more abstract type of intonation grammar is the auto-segmental intonation grammar. This type of grammar also concentrates on changes in F_0, but proposes a more abstract description such as high tones (indicated with H) or low tones (indicated with L). If a tone (H, L) is accent lending (prominent) it is indicated with an asterisk (*). The actually realized intonation contour is not described. Human labeling according to this grammar shows rather low consistency rates. Maximally 56% correspondence for H* and L*+H was reported in Reyelt (1995). See also Syrdal et al. (2001).
For speech technology purposes, human pitch accent labeling is not consistent enough, and does not contain variation about the prominence of these accents. Wightman & Ross (1999) and Wightman et al. (2000) suggested using a robust variant of TOBI. This variant is called TOBI lite. Attempts to derive an automatically encoded TOBI intonation contour are presented in Véronis & Campione (1998) and Tournemire (1998). Their results are unclear.

It is felt in the present study that the above described research concentrate too much on F_0 changes. The perception of prominence is not exclusively related to smaller or larger F_0 changes. Supportive evidence for this is derived from a small listening experiment done with students at the Institute of Phonetic Sciences (reported in Streefkerk et al., 1997). In this pilot experiment 30 natural sentences were presented to listeners who were asked to mark the words spoken with emphasis. These sentences were presented in a normal version and in a version where the F_0 was monotonized via the PSOLA technique. The duration variations, as well as the variations in intensity and spectral quality were still present. All eight participants appeared to be able to mark prominence even in these monotonous sentences. Even seven or eight of them uniformly marked not less than six words as being prominent under both conditions. The correspondence matrix of the sentences, presented with and without pitch movements, is given in Appendix table A.2.1. The task is of course more difficult in the monotonized sentences, but listeners were still able to mark prominence and even sometimes achieved a unanimous judgment. It can be concluded that prominence is not only evoked by pitch movements, but that other acoustic correlates, such as duration and loudness, are likely to be additional cues for the listener to perceive prominence.

2.1.2 Our approach

We define 'prominence' operationally by asking listeners to mark the perceptually 'outstanding' parts in speech. The main question is how to design a listening experiment to mark these more prominent parts of speech.

This question can be subdivided into the following sub-questions:

- Should words / syllables be marked for prominence by experts or by naive listeners?
- Is a majority judgment possible and necessary?
- Should all words / syllables be marked for prominence?
- Is a binary scale useful for marking prominence?
- What is the unit to be marked: word group, word or syllable?

All these questions must be answered in order to find a proper operational definition of prominence.
In our study naive listeners instead of experts were chosen as labelers for marking prominence. Users of a language do use the differences in prominence in their daily conversation, but they are not aware of it most of the time.
To make sure that this research is not based on the possible coincidental prominence marks of only one listener, a group of listeners participated in the labeling experiment. Another advantage of taking more than one listener is that the agreement of the listeners can be calculated, which gives more insight in the consistency between listeners. The labeling task is much easier if listeners were asked to mark only every individual word or syllable spoken with emphasis, instead of asking them to mark all the words of an utterance for relative prominence, as done by Portele & Heuft (1997) and Grover et al. (1997). It was decided to make the judgments of the individual listeners binary (a word or syllable could either be prominent or not), which makes the task easier. By taking the sum of all individual marks a gradient prominence scale becomes available.

There are also other pragmatic issues that have to be taken into account. On the one hand, in order to increase the validity, the task must be easily interpretable for naive listeners and we aim at as little guidance to the listener as possible, because the more instructions are given the more influential is the interpretation of the researcher, and we want not to investigate our ideas of what prominence is, or should be, but those of naive listeners. This is often translated in ideas of how many word / syllables have to be prominent in a sentence. On the other hand with less guidance one has to be careful, to make listeners understand the task properly.
The method must be useful for a large speech corpus and listeners must be able to mark prominence on-line.

For Text-to-Speech (TTS) and automatic speech recognition (ASR) systems, prominence labeling on word level is sufficient to be useful in applications. Text-to-Speech systems do not need a labeling at the syllable level, because the lexicon gives information on the position of lexical stress. Word prominence information is acoustically realized by using the syllable information in the lexicon. So one does not need to mark individual syllables, it goes via the information in the lexicon.
For detection with acoustical information one might argue that prominence can only be acoustically detected by looking for the syllable that is the most outstanding one,

which does not have to be the lexically stressed one. So, one might argue that prominence detection must be trained and tested on syllables. In ASR a lexicon is used as well, and in ASR words go into competition with each other; syllables themselves do not play a role. Via this lexicon the lexically stressed syllable could be estimated, if marked. If word stress is not marked, ASR could not distinguish between words such as *AchterRUIT (rear-window)* and *ACHTeruit (in reverse)* even if one knows the prominence of the individual syllables in the training. So, for ASR prominence labeling on words is enough, unless one does not want to re-rank the individual syllable scores with the prosodic information from the lexicon.

In the next section, we will describe the speech material used.

2.2 Speech material

In our research we have chosen to use the spoken sentences from the Dutch Polyphone Corpus (Damhuis et al., 1994), because, firstly, this telephone speech in the Dutch Polyphone Corpus is more realistic for various speech technology applications than elicited speech recorded in an anechoic room. Secondly, using the Dutch Polyphone Corpus gives us access to a large speech database, which is easy to use and is readily available. The original recordings were made by KPN and SPEX, who made this corpus available on CD-ROM for academic purposes. A brief description of the Dutch Polyphone Corpus follows in the next subsection.
The choice of the speech material used has large consequences for the results achieved and for the conclusions that can be drawn. The speech material used in this study imposes several limitations. This material is especially designed for speech recognition.
The specifications of any corpus concern the speakers, the environment, type of speech, speaking rate, dialectical background, gender, age and socio-economic status. The acoustical surroundings of the recordings can also differ a lot. For instance in an anechoic room no background noise is recorded, but at home and certainly in a car a lot of background noise might be added to the speech signal. To improve all-purpose speech recognition one has to cope with all these and other factors, so we consider it advantageous for speech recognition to have a number of these variations in our Polyphone speech material.
However, for speech synthesis the most important goal generally is to make a natural and pleasant voice. All the speaker and surrounding variation is not very useful for speech synthesis, but variations in speaking styles and genres, such as free conversation, monologue, dialogue, retold, and read-aloud, certainly are. All this introduces variation in the speech material and has its consequences for the prosody, which is more important for speech synthesis than speaker variation is. Speech synthesis actually needs one good voice showing sufficient variation, especially in longer text. Unfortunately, such variation in different speaking styles is not present in the speech material used by us. Only read-aloud sentences are available.
Improving TTS intonation can thus only be a secondary aspect of the present research. No contextual information can be derived from these sentences, since they were all separately spoken and unrelated. Moreover this typical type of speaking

style may cause a typical intonation. It should thus be clear that the design of the present database is not ideally suited to improve Text-to-Speech systems.

2.2.1 Recordings

The Polyphone Corpus was recorded over the telephone; in this way the speaker is in his familiar environment, but background noises are then unavoidable. The recordings were sampled with a frequency of 8 kHz in 8 bit A-law coded samples and were stored on CD-ROM. Each record (file) was provided with a NIST header. An example is given in Appendix A 2.2. The header fields concern the record and give specific speaker information. Extra information was added in a post-processing cycle, providing a word-by-word transliteration, a transliteration of extra sounds such as noisy breathing, demographic data, and the quality of each recorded item. The assessment could be 'OK', 'noise', 'garbage' or 'other'. About 97% of the recordings are of good quality ('OK').

2.2.2 Material

For each speaker 32 different speech items were recorded, which vary from digits and spelling out names to answering questions given on the instruction paper. For the present project only the sentences that were read-aloud were used (five different sentences per speaker). These sentences are constructed in such a way as to be phonetically rich, which means that all sounds of the Dutch language system occur at least once per set of five sentences. In order to meet this requirement an electronic version of a newspaper was scanned, and the required sentences were selected. This resulted in 12,500 different newspaper sentences, not shorter than four words and not longer than 80 characters. The grammatical structure of these sentences is simple (most of the time declarative sentences with only one main clause and no nested sentences). Questions are particularly rare. Twice two different speakers speak every sentence of the set of 12,500. Approximately 5000 speakers read five sentences each.

2.2.3 Speakers

The speakers recorded in the Dutch Polyphone Corpus are of various ages, and come from different regions of the Netherlands. This introduces accents from different dialectal backgrounds. The speakers themselves specified from which province they came (Groningen, Friesland, Drente, Overijssel, Flevoland, Gelderland, Utrecht, Noord-Holland, Zuid-Holland, Zeeland, Noord-Brabant, Limburg). The distribution of the speakers coincides more or less with the distribution of the population. Their ages ranged from 16 to 80. The socio-economic status is specified by the education level, which ranged from elementary school, to secondary school and college / university level.

2.2.4 Selected speech material

In our research a total of 2244 sentences from this Polyphone Corpus were used. Only the two pilot experiments as described in section 2.3 use another randomly selected set of 81 sentences from this corpus. These sentences are excluded from the general description of the speech material used. A set of 1244 sentences served as a so-called training set and another 1000 as an independent test set. The 1244 sentences of the training set were used for lexical / syntactic analysis (chapter 3), and for acoustic analysis (chapter 4) and this same set was also applied to train neural networks for prominence classification (chapter 5). The test set of 1000 sentences was applied in order to test the prediction of linguistic and acoustic features. Using different test and training sets makes the results more reliable and controllable, so that the rules predicting prominence from textual information and a prediction with neural networks will not appear to be over-specified, because of the use of only a limited set of sentences.
The sentences used were semi-randomly chosen. Based on the quality of the sound recordings not always all five sentences spoken per speaker were useful: sometimes only two sentences per speaker were chosen. In total the training set plus the test set contain sentences from 497 different speakers.
The demographic data are specified in the Appendix in tables A 2.3 and A.2.4.

In order to investigate what the most appropriate unit (word or syllable) is to mark prominence, two pilot experiments with the same set of 81 sentences were conducted. The details of the two pilot experiments are reported in Streefkerk et al. (1997), but a brief summary is presented in the next subsections.

2.3 Two pilot experiments to define prominence

For an operational definition of prominence one of the remaining questions is: What is the appropriate unit for marking prominence (word or syllable). In the case of word groups the listener can mark adjacent words or syllables, so actually the question that remains is: should listeners mark words or syllables? In order to learn about the differences on marking behavior, one experiment on syllable prominence marking and one experiment on word prominence marking were conducted. The criterion for choosing between the two approaches is based on the amount of information obtained, i.e. on the validity and on the consistency of the responses.

2.3.1 Method

For both experiments the acoustical presentation of the sentences was identical, only the instruction and the display of the text on the monitor differed. For the word-marking experiment the listeners saw a normal orthographic text on the monitor, and got the instruction to mark all words spoken with emphasis (in the Dutch instruction *nadruk,* see Appendix A 2.5). For the syllable-marking experiment a white space was displayed between syllables, plus a hyphen for syllables belonging to one word. An example is given next. The literal translation is: *The short-lived rise of the dollar is over.*

De	*kort-*	*ston-*	*di-*	*ge*	*op-*	*mars*	*van*	*de*	*dol-*	*lar*	*is*	*voor-*	*bij.*
o	o	o	o	o	o	o	o	o	o	o	o	o	o

In the syllable experiment, the task was to mark all syllables spoken with emphasis. For each type of experiment the 81 sentences were presented to 8 listeners.

2.3.2 Results

An example of the individual and added word marks for one specific sentence is given in table 2.1. These cumulative marks can be interpreted as a gradient prominence scale. If a word or a syllable is marked by all of the listeners (in these pilot experiments there were eight listeners) a word or syllable is apparently highly prominent. Similarly, if no or only one listener marks a word, this word is apparently less prominent. The cumulative marks are also an indication of consistency. If there are a lot of words / syllables which are only marked by one listener, the between subject agreement is not high. Table 2.2 shows the cumulative prominence marks (absolute and relative) for the syllable experiment and for the word experiment.
The total number of judgments in the two listening experiments differed. Per listener the mean number of prominence judgments per sentence in the word-perception experiment is 2.9 (0.7 Std. Dev). This is substantially lower than in the syllable-marking experiment (5.1, 1.7 Std. Dev). A t-test for two samples clearly shows that these two means differ significantly ($t = -3.356$, $\nu = 14$, $p \leq 0.005$).

Table 2.1: A part of the raw data matrix of the word perception experiment. The individuals and the cumulative word prominence labeling over all eight listeners are given. The literal translation of this example sentence is: *There goes a bus at half past two from Amsterdam to Utrecht.*

	Listener								
Word	1	2	3	4	5	6	7	8	Total
Er	0	0	0	0	0	0	0	0	0
gaat	1	1	0	1	1	1	1	1	7
om	0	0	0	0	0	0	0	0	0
half	0	1	1	0	0	1	0	0	3
drie	0	0	1	0	0	0	0	0	1
een	0	0	0	0	0	0	0	0	0
bus	1	0	0	0	1	1	1	1	5
uit	0	0	0	0	0	0	0	0	0
Amsterdam	1	1	1	1	1	1	0	1	7
naar	0	0	0	0	0	0	0	0	0
Utrecht	1	1	1	1	1	1	1	1	8
Total per listener	4	4	4	3	4	5	3	4	31
Mean per sentence								31 / 8 = 3.9	

Table 2.2: For both the syllable and the word listening experiment the cumulated prominence marks, and the percentage values, are shown. For example in column 'Freq. (Syllable Experiment) the number 77 means that 77 syllables have been marked by all eight listeners as being prominent.

	Syllable Experiment			Word Experiment		
Cumulative marks	Freq.	Cum. num of marks	%	Freq.	Cum. num of marks	%
0	519	-	35.5	432	-	50.6
1	296	296	20.3	66	66	7.7
2	162	324	11.1	41	82	4.8
3	90	270	6.2	53	159	6.2
4	63	252	4.3	44	176	5.2
5	61	305	4.2	52	260	6.1
6	92	552	6.3	61	366	7.2
7	100	700	6.8	51	357	6.0
8	77	616	5.3	53	424	6.2
Total	1460	3315	100	853	1890	100
Mean per sentence per listener		5.1			2.9	

If the listeners had marked the prominent words / syllables in completely random fashion, this frequency table would have looked different. The total number of syllable marks of the eight listeners is 3315; this number divided by the total number of syllables (1460) produces 2.3. Thus, if the eight listeners had marked the syllables randomly, each syllable would on average have received a cumulative mark of 2.3. Similarly, such a random order value of cumulative marks can be calculated for the word experiment. The total of 1890 marks divided by 853 results in 2.2. The example sentence in table 2.1 then would have shown a maximum cumulative mark of approximately 2.

The distributions given in table 2.2 are most certainly not a random distribution of prominence marks. However, there are differences between the two types of experiments. The syllable experiment shows relatively high numbers of marks given by only one listener compared to the word experiment (20.3% versus 7.7%), and the scores for just two listeners also show a difference (11.1% versus 4.8%). The relative numbers of values, concerning the majority of the listeners marking a given syllable / word as prominent, do not show such differences.

The agreement between- and within-listeners can be measured from crosstables by using Cohen's Kappa κ (Cohen, 1960). This agreement measure calculates the agreement corrected for chance agreement. The formula is as follows:

$$\kappa = \frac{N \sum_{i=1}^{r} F_{ii} - \sum_{i=1}^{r} R_i C_i}{N^2 - \sum_{i=1}^{r} R_i C_i}$$

where N is the total number of items and r the number of rows in the table. F_{ii} is the number found in the cells of the diagonal, R_i is the row sum and C_i the column sum.
To get a feeling for this agreement measure κ, the following theoretical examples have been constructed, which will explain the behavior of this agreement measure. Cohen's Kappa (κ) lies between -1 and 1, where '1' means total agreement and '0' means no agreement beyond chance level. If always-opposite judgments were given a Cohen's Kappa of -1 would result as in example e).
Some examples are given to show the effect of asymmetrically distributed data. Our listener judgments show such asymmetry. Examples a), c) and e) show a symmetrical distribution of the data; the column sums and the row sums are the same. This is different in example b) and d). In these two examples the distribution is not symmetrical; the first column adds up to 300 whereas the second column adds up to 100. For the first two examples a) and b) $\kappa = 0$: there is only agreement at chance level. For the evenly distributed table a) this is clear, but it is also the case for the not evenly distributed table b). In tables c) and d) the data are distributed in such a way that the highest possible agreement is given, for table c) $\kappa = 1$, which means total agreement. However, for table d) $\kappa = 0.5$, which is the highest possible agreement for this distribution as on the one side only 100 times a one is given whereas on the other side 200 times a one is given. This means that there is a maximum overlap of 100 times a one. So in table d) the agreement measure is rather low ($\kappa = 0.5$), but a higher value is not possible for this kind of distribution. For the last example e) Cohen's Kappa is -1; this happens when only opposite judgments are given.

a)

	0	1	Total
0	100	100	200
1	100	100	200
Total	200	200	400

$\kappa = 0$

b)

	0	1	Total
0	150	50	200
1	150	50	200
Total	300	100	400

$\kappa = 0$

c)

	0	1	Total
0	200	0	200
1	0	200	200
Total	200	200	400

$\kappa = 1$

d)

	0	1	Total
0	200	0	200
1	100	100	200
Total	300	100	400

$\kappa = 0.5$

e)

	0	1	Total
0	0	200	200
1	200	0	200
Total	200	200	400

$\kappa = -1$

The actual data of the two listening experiments are also not very symmetrically distributed: some listeners have marked substantially more words / syllables as being prominent than others. A crosstable of two listeners (listener one as rows and listener six as columns) is presented in table 2.3 below.

Table 2.3: Crosstable of listener one and six, who mark prominent words.

	Listener 6		
Listener 1	Non-prom	Prom	Total
Non-prom	459	115	574
Prom	41	238	279
Total	500	353	853

82% $\kappa = 0.61$

For this crosstable Cohen's Kappa is 0.61, which is rather high taking into account the fact that this table is similarly distributed as in example d). Listener one marked only 279 of the 853 words as prominent whereas listener six marked 353 words as prominent, which shows that listeners use different thresholds to judge what is prominent. These two listeners cannot agree completely, because they do not have the same number of prominent marks. The agreement between listener one and listener six is expressed in Cohen's Kappa ($\kappa = 0.61$), which becomes one entry in table 2.4.

Table 2.4: Agreement (Cohen's Kappa) between eight listeners in the word experiment.

Listener	1	2	3	4	5	6	7	8
1	-	0.59	0.58	0.41	0.54	0.61	0.55	0.50
2	-	-	0.59	0.55	0.89	0.69	0.53	0.53
3	-	-	-	0.48	0.59	0.63	0.60	0.58
4	-	-	-	-	0.53	0.47	0.46	0.54
5	-	-	-	-	-	0.59	0.54	0.53
6	-	-	-	-	-	-	0.48	0.53
7	-	-	-	-	-	-	-	0.51
8	-	-	-	-	-	-	-	-
Mean Cohen's Kappa								0.56

Table 2.5: Agreement (Cohen's Kappa) between the eight listeners in the syllable experiment.

Listener	1	2	3	4	5	6	7	8
1	-	0.70	0.60	0.34	0.66	0.45	0.67	0.22
2	-	-	0.64	0.50	0.69	0.48	0.84	0.21
3	-	-	-	0.50	0.63	0.53	0.67	0.16
4	-	-	-	-	0.39	0.45	0.39	0.13
5	-	-	-	-	-	0.45	0.71	0.20
6	-	-	-	-	-	-	0.52	0.39
7	-	-	-	-	-	-	-	0.20
8	-	-	-	-	-	-	-	-
Mean Cohen's Kappa								0.48

For each paired combination of listeners for the word experiment and for the syllable experiment, the Cohen's Kappa values are given in table 2.4 and 2.5, respectively. For the word experiment Cohen's Kappa is smaller than 0.45 only once; for the syllable experiment Cohen's Kappa is ten times smaller than 0.45. It is apparent from table 2.5 that listener eight is not a very useful listener. Perhaps this listener did not understand the task sufficiently. The listeners in the word experiment reach an agreement of $\kappa \geq 0.55$ 13 times. For the syllable experiment this happens only ten times. The results for the syllable experiment indicate less agreement between the listeners than for the word experiment.

2.3.3 Conclusion

The two different prominence-marking methods show differences in the number of marked words versus marked syllables. These two different methods do not result in the same prominence marking distribution. Prominence marking on syllables results in a more detailed prominence distribution of the sentences than on words.
A priori it is difficult to decide which method is better. However, it appears that the syllable method seems to show less agreement between listeners. The differences in the agreement between the listeners are not so striking, but we have to choose between these two methods of marking prominence. The word method is chosen because, firstly, a more detailed marking in terms of prominent syllables is apparently not needed and, secondly prominence marking on words is more valid. Words are more meaningful units for naive native listeners than syllables. They can mark words online for prominence using bottom-up and top-down information. Finally, marking word prominence easier to perform than marking syllables and could therefore be less time consuming.

2.4 Main experiment on assigning prominence

2.4.1 First set marked by ten listeners

A prominence labelling experiment was carried out in the following way. The first set of 1244 sentences was divided into two subsets: one with 500 sentences spoken by 100 speakers, and another set of 744 sentences spoken by 173 speakers. To test the within-listener consistency, the first 50 sentences of the first set were presented twice to each listener. The 550 sentences (500 + 50) of the first listening set, as well as the 744 sentences of the second listening set were each presented over four sessions, of approximately one hour each. The sentences from both listening sets were presented to ten listeners. Each group of ten listeners was not necessarily composed of the same participants for each set: only a few participated in both sets. The sentences were presented individually in random order under computer control to compensate for possible learning effects. The listeners were students from the Humanities Faculty of the University of Amsterdam and they were paid for doing this task. The listeners received no training. The experiment was controlled by a UNIX workstation and was designed in such a way that the written words of each sentence were displayed on the monitor with a button underneath each word. The monitor screen looked like this:

Vaak	meet	men	aan	de	inhoud	van	de	kassa	het	welslagen	van	een	project	af.
o	o	o	o	o	o	o	o	o	o	o	o	o	o	o

A translation of this sentence is: *Often one determines a project's success by the amount of money in the cashbox*. The listeners could click on the buttons corresponding to words perceived as being spoken with emphasis (*nadruk*, for the

whole instruction in Dutch see Appendix A 2.5). While the sentence was displayed on the monitor, the spoken version was presented to the subject through headphones. Each sentence was presented a maximum of three times. To move on to the next sentence participants could click on a button labelled *'klaar' (finished)*, which means that they did not need to listen to unnecessary repetitions of one sentence. All the subjects did this task without any complications. The results were stored on disk for further processing. The results of the listening experiment will be discussed in the following sections. First, the general results are presented, after which differences between and within listeners will be explained in detail.

2.4.1.1 Resulting prominence marks

In table 2.6 the results of both subsets (500 and 744 sentences) on the prominence-marking task are presented together. The number of marks per word has been added. This table 2.6 presents the number of words found with a given number of cumulative prominence marks. The second column (Freq.) gives absolute numbers: for example the number 802 means that this number of words have been marked as being prominent by eight out of ten listeners. The third column gives the percentages. The 50 sentences presented twice were included only once in these data (only the marks of the sentences which were presented the second time to the listener were included). It is worth pointing out that approximately half of the words (45.5%) were not marked as being prominent by any of the listeners; one or more of the listeners marked the remaining 54.5% of the words as prominent. 8.9% of the words were only marked by one listener and everybody agreed about 4% of the words as being prominent.

Table 2.6: The absolute and the relative numbers of the cumulative prominence marks.

Cumulative marks	Freq.	%
0	5950	45.5
1	1162	8.9
2	679	5.2
3	595	4.5
4	586	4.5
5	603	4.6
6	645	4.9
7	813	6.2
8	802	6.1
9	739	5.6
10	518	4.0
Total	13092	100

The prominence marks can be used in several ways. Prominent and non-prominent words can be defined by, for instance, a majority judgment of the listeners. The prominence judgments can also be interpreted as a prominence scale from 0 to 10. Since there is variability between listeners, it cannot be expected that the scale from 0 to 10 will represent eleven separate classes. To get a more meaningful prominence label a hierarchical cluster analysis will be used. Which prominence values will be clustered and how consistent the listeners were, is the topic of the next two sections.

2.4.1.2 Differences within and between listeners

Between-listener agreement shows reliability, whereas within-listener agreement shows consistency. The differences within and between listeners can be studied via those 50 sentences that are presented twice to the ten listeners. Table 2.7 presents the total number of prominence judgments over all 50 sentences per listener.

Table 2.7: Number of prominence judgments per listener after first and second listening.

Listener	1	2	3	4	5	6	7	8	9	10	Total
First	71	50	160	165	135	132	50	109	156	172	1200
Second	71	51	165	202	130	211	50	149	209	158	1396

The total number varies from 50 to 211 for the 50 sentences that were judged twice by each listener. There are listeners who marked many words as being spoken with emphasis, as well as one listener who marked only one word per sentence as prominent (listener seven). Furthermore, listeners can be differentiated by the consistency of their marking of words both times they listened: those who marked approximately the same number of words each time (listener one, two, three, five, seven and ten), and those who marked more words the second time (listener four, six, eight and nine).

Table 2.7 gives thus rise to the following questions. Although some listeners judged many more words as prominent than others, does the larger set of words include the words in the smaller set? If the set of prominently marked words is a completely different set, then the reliability is poor. If, however, there is substantial overlap and one listener just marks additional words, then there is only a difference in threshold. To investigate this, a number of crosstables were constructed. Table 2.8 gives the scores of listener one and listener three who marked nearly the same number of words both times. So these tables show the individual listener consistency. Of the 71 prominence judgments of listener one, 48 are the same in both sessions. The agreement is $\kappa = 0.63$. The consistency of listener three is much better ($\kappa = 0.81$) 142 prominence marks agree in both sessions.

Table 2.8: Crosstables of the first and second listening session (of the 50 sentences presented twice) of listener one and three. These tables show individual listener consistency.

	Listener 1 second				
Listener 1 first	Non-prom	Prom	Total		
Non-prom	425	23	448		
Prom	23	48	71		
Total	448	71	519	91%	$\kappa = 0.63$

	Listener 3 second				
Listener 3 first	Non-prom	Prom	Total		
Non-prom	336	23	359		
Prom	18	142	160		
Total	354	165	519	92%	$\kappa = 0.81$

Another question is, up to what level do the judgments of listener one, who is very economical with his judgments, correspond with those of listener three, who gives many more judgments.

Table 2.9 shows such a between-subject comparison between listener one and listener three. The 71 words marked in the 50 sentences of listener one agree for the greater part (62, 69, 65, 70) with the 165 or 160 of listener three for the first time / first time and second time / second time comparison, as well as for the first time / second time and second time / first time comparison. Listener three marked many more words as prominent, which may be so because this listener has a lower threshold for perceiving prominence than listener one. The overall agreement between these two listeners can be expressed in Cohen's Kappa, which we have already introduced in section 2.3.2. The agreement values for these two listeners (one and three) are also given in table 2.9. The agreement value of listener one and three is only approximately $\kappa = 0.5$, mainly because of the difference in total marks of these two listeners.

Table 2.9: Crosstable comparison of listener one and three, which shows between-listener differences.

Listener 3 first

Listener 1 first	Non-prom	Prom	Total		
Non-prom	350	98	448		
Prom	9	62	71		
Total	359	160	519	79%	$\kappa = 0.43$

Listener 3 second

Listener 1 second	Non-prom	Prom	Total		
Non-prom	352	96	448		
Prom	2	69	71		
Total	354	165	519	81%	$\kappa = 0.49$

Listener 3 first

Listener 1 second	Non-prom	Prom	Total		
Non-prom	353	95	448		
Prom	6	65	71		
Total	359	160	519	81%	$\kappa = 0.46$

Listener 3 second

Listener 1 first	Non-prom	Prom	Total		
Non-prom	353	95	448		
Prom	1	70	71		
Total	354	165	519	82%	$\kappa = 0.50$

Table 2.10: Cohen's Kappa for all combinations of listeners for the first-first combination (above the diagonal) and second-second combination (below the diagonal).

Second-second \ First-first	1	2	3	4	5	6	7	8	9	10
1	-	0.34	0.43	0.43	0.47	0.44	0.39	0.39	0.41	0.41
2	0.45	-	0.35	0.34	0.37	0.37	0.49	0.36	0.32	0.26
3	0.49	0.36	-	0.72	0.72	0.67	0.36	0.48	0.75	0.62
4	0.40	0.28	0.75	-	0.65	0.57	0.32	0.52	0.67	0.60
5	0.46	0.35	0.68	0.61	-	0.63	0.63	0.55	0.69	0.63
6	0.31	0.22	0.66	0.72	0.55	-	0.34	0.55	0.63	0.58
7	0.32	0.42	0.36	0.27	0.39	0.24	-	0.27	0.31	0.21
8	0.40	0.32	0.65	0.65	0.57	0.66	0.30	-	0.51	0.46
9	0.35	0.26	0.73	0.80	0.54	0.80	0.25	0.63	-	0.61
10	0.41	0.31	0.69	0.65	0.66	0.59	0.32	0.63	0.65	-

Mean $\kappa = 0.50$

Table 2.11: Cohen's Kappa for all combinations of listeners for the second-first combination (above the diagonal) and first-second combination (below the diagonal). On the diagonal the within-listener agreement is given.

First-second \ Second-first	1	2	3	4	5	6	7	8	9	10
1	**0.63**	0.34	0.46	0.45	0.42	0.43	0.34	0.31	0.45	0.37
2	0.56	**0.46**	0.38	0.34	0.37	0.37	0.55	0.32	0.34	0.28
3	0.50	0.36	**0.82**	0.72	0.71	0.67	0.35	0.52	0.78	0.64
4	0.37	0.27	0.76	**0.69**	0.65	0.59	0.27	0.48	0.71	0.64
5	0.49	0.37	0.70	0.64	**0.76**	0.60	0.37	0.54	0.69	0.60
6	0.33	0.22	0.67	0.56	0.58	**0.65**	0.22	0.49	0.67	0.60
7	0.34	0.56	0.36	0.34	0.39	0.40	**0.58**	0.36	0.36	0.24
8	0.42	0.31	0.68	0.61	0.61	0.63	0.29	**0.57**	0.65	0.58
9	0.35	0.22	0.72	0.62	0.58	0.61	0.26	0.46	**0.73**	0.64
10	0.45	0.32	0.66	0.64	0.66	0.64	0.27	0.53	0.64	**0.73**

Mean $\kappa = 0.50$

In table 2.10 and table 2.11 Cohen's Kappa is presented for each possible combination of listeners, and for first and second presentation.

For instance, the first time listeners one and three marked the 50 sentences for prominence the agreement measure $\kappa = 0.43$, which appears in table 2.10 above the diagonal.
The agreement within listeners is given in the shaded cells on the diagonal of table 2.11. Listeners three, five, nine and ten were very consistent ($\kappa > 0.7$) whereas listener two was less consistent ($\kappa < 0.5$).
The agreement between listeners is high for listeners three, four, eight, nine, and ten in all combinations. There are also listeners who agree less, such as listeners two and seven. Not only are they inconsistent within their own judgments, they also do not agree with other listeners. But it must be mentioned that these two listeners actually have a very low number of total judgments: they only give a total of 50 marks (table 2.7). Agreement between the listeners is not as low as found for one listener in the pilot experiment (see section 2.3.2 table 2.5).
Differences within and between listeners exist, but we consider the accumulated prominence marks per word to be a useful alternative for a gradient prominence labeling of the speech material.

These cumulative prominence marks can be clustered in various ways, for instance four or two groups, to reduce the labeling variation. This is the topic of the next section.

2.4.1.3 Clustering of the cumulative prominence marks

The cumulative prominence marks for each word form a prominence scale. If only one listener marks a given word (resulting in a score of 1) this word is considered less prominent than if all ten listeners mark this word (resulting in a score of 10). Since in this study ten listeners could mark a specific word, this results in an 11-point prominence scale (from 0 to 10). An 11-point scale may be too fine-tuned whereas it is also directly related to the (rather arbitrary) number of subjects we had: if we had asked only eight listeners to mark the sentences the prominence scale would have been a 9-point scale from 0 to 8. Putting similar groups together by means of hierarchical cluster analyses reduces this scale to a more robust one. To this end, we constructed a confusion matrix for the twice-judged 50 sentences only (table 2.12). This matrix shows that some points of the prominence scale are more confused than others and therefore it is justified to put them together. Based on this confusion matrix a similarity matrix was calculated and a cluster analysis was carried out.

Statistically the similarity of points on a scale can be expressed in different ways, which results in different cluster methods with differing results. Here the hierarchical cluster scheme of Johnson was chosen (Johnson, 1967). This cluster method is based on matrix manipulations. This hierarchical cluster method uses a similarity matrix and has the advantage that each point does not have to be initially

Table 2.12: Confusion matrix on the basis of the 50 sentences that have been marked twice by 10 listeners. The entry 20 in cell (0;1) means that it occurred 20 times that words were marked for prominence just once in the first listening session, whereas in the second session these words were not marked at all. The gray scaling is explained in the text.

Resulting prominence scale from the first listening session

Resulting prominence scale from the second listening session	0	1	2	3	4	5	6	7	8	9	10	Total
0	244	20	4	1	3	0	0	0	0	0	0	272
1	10	6	9	10	4	1	0	0	0	0	0	40
2	4	4	6	5	7	2	3	0	0	0	0	31
3	0	0	2	6	6	5	2	1	0	0	0	22
4	0	0	0	2	6	7	3	3	0	0	0	21
5	0	0	1	0	3	7	11	7	2	0	0	31
6	0	0	0	0	0	2	3	4	5	2	1	17
7	0	0	0	0	0	1	4	11	5	4	0	25
8	0	0	0	0	0	0	2	8	11	10	1	32
9	0	0	0	0	0	0	0	1	6	6	7	20
10	0	0	0	0	0	0	0	0	0	4	4	8
Total	258	30	22	24	29	25	28	35	29	26	13	519

represented as a point in a Euclidean space. Before this cluster analysis can be done, the confusion matrix given in table 2.12 must be made symmetric and must be scaled for the different row and column totals, which results in table 2.13. The numbers on the diagonal are given as 0, so they no longer play a role for the clustering. Each cell is divided by the mean of its row and column total in order to correct for the different totals. To make the matrix symmetric around the main diagonal, each cell is the average of its mirrored counterpart. For instance, the new value in cell (2;5) and cell (5;2) is the average of the old values of these cells. Higher values on the prominence scale represent more similarity. The cells that are more alike are exactly the points on the prominence scale which are (relative to the row and column totals) likely to be more confused. With the cluster analysis according to Johnson (Maximum method), these points were clustered. Different scaling techniques and methods (Maximum, Minimum, Mean) were tried, but the scaling technique described above (with the maximum method) yields in the most meaningful and interpretable result.

Table 2.13: Similarity matrix made from the confusion matrix given in table 2.12. Each cell is divided by the mean of its row and column total. The matrix is made symmetric by taking the mean of the mirrored cells around the main diagonal.

Prominence scale	Prominence scale 0	1	2	3	4	5	6	7	8	9	10
0	0	0.1	0.03	0	0.01	0	0	0	0	0	0
1	0.1	0	0.21	0.16	0.06	0.02	0	0	0	0	0
2	0.03	0.21	0	0.14	0.12	0.04	0.05	0	0	0	0
3	0	0.16	0.14	0	0.16	0.11	0.04	0.02	0	0	0
4	0.01	0.06	0.12	0.16	0	0.2	0.06	0.05	0	0	0
5	0	0.02	0.04	0.11	0.2	0	0.23	0.13	0.03	0	0
6	0	0	0.05	0.04	0.06	0.23	0	0.15	0.14	0.05	0.03
7	0	0	0	0.02	0.05	0.13	0.15	0	0.21	0.1	0
8	0	0	0	0	0	0.03	0.14	0.21	0	0.29	0.02
9	0	0	0	0	0	0	0.05	0.1	0.29	0	0.33
10	0	0	0	0	0	0	0.03	0	0.02	0.33	0

The results of various clustering options from five to two clusters are given in table 2.14. If the 11-point scale is reduced to five categories, zero stands for one category, one and two are put together in one cluster, three is a cluster on its own, whereas four, five and six, as well as seven, eight, nine and ten, form two more clusters. These five clusters could be further reduced to four, three or two clusters, as shown in table 2.14. In figure 2.1, for the clustering as shown in table 2.14, a corresponding dendrogram is painted. The dendrogram shows how the different clusters originated, and on which distance the different prominence values are gathered. The different values and its clustering are given on the vertical. Firstly, nine and ten are clustered, then eight is added to this cluster, and thirdly five and six are clustered, and so on. Which distances these clusters have, can be seen in the little boxes. Using the four-cluster scale is a good alternative, as no groups stand alone except the zero, and the original scale is agglomerated enough (the dotted line in the dendrogram), while leaving us with enough variation. These four clusters are: 0 representing no prominence, 1-2 and 3-6 representing the in-between categories and 7-10 representing the most prominent words. The points that are put together on the prominence scale are indicated in grey cells in the original confusion matrix in table 2.12.

Table 2.14: Resulting number of clusters (from five to two) of the hierarchical cluster analyses (Maximum method). The roman numbers are used for the new prominence classes.

Prominence scale	5 Clusters	4 Clusters		3 Clusters	2 Clusters
0	1	1	**0**	1	1
1	2	2	**I**	2	2
2	2	2		2	2
3	3	3		2	2
4	4	3	**II**	2	2
5	4	3		2	2
6	4	3		2	2
7	5	4		3	2
8	5	4	**III**	3	2
9	5	4		3	2
10	5	4		3	2

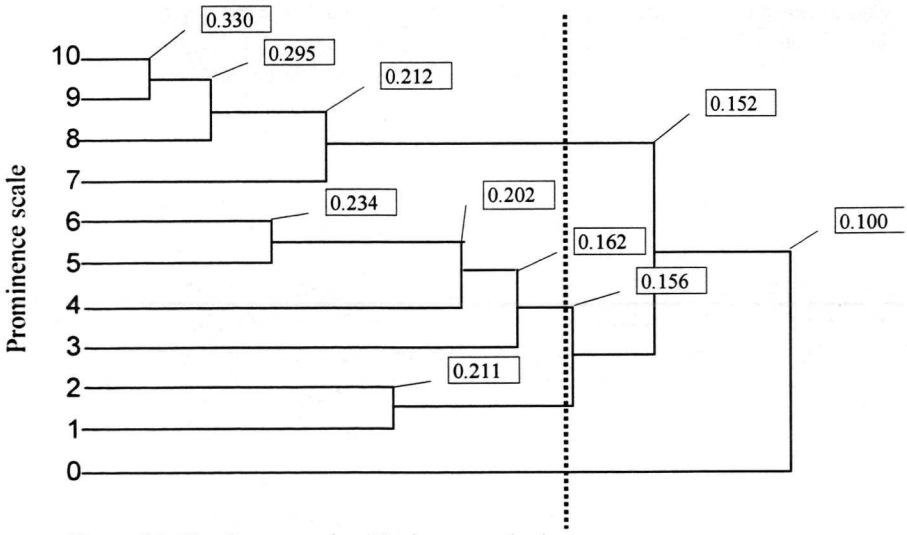


Figure 2.1: Dendrogram using Maximum method.

Table 2.15: The absolute and relative numbers of words each of four prominence classes, which is a result of the agglomerated prominence scale.

Prominence class	Freq.	%
0	5950	45.4
I	1841	14.1
II	2429	18.6
III	2872	21.9
Total	13092	100

The relative and the absolute numbers of words in a given prominence class are presented in table 2.15. These clustered numbers are those used for further analysis in chapter 3, chapter 4 and chapter 5.

2.4.2 Second set marked by one 'optimal' listener

As we will see in chapters 3, 4 and 5, the prominence prediction and classification methods were developed with the help of the training set; in order to properly test these methods an independent test set is required.
The labeling of the independent test set was performed in a slightly different way; mainly for efficiency reasons. The criteria for selecting the test sentences are completely the same as those for the training material. Simply put, the next acceptable 1000 sentences of the Dutch Polyphone Corpus have been taken (more details can be found in section 2.2.4 of this chapter).
The prominence marks of the ten listeners are not required for this testing purpose. For the development of the classification methods a majority judgment is used. This results in a prominent / non-prominent distinction. A useful alternative for asking ten listeners to mark the whole set of 1000 sentences would be to ask only one representative listener. This listener should preferably have the highest agreement with the other listeners and should be consistent in his / her judgments. The results in tables 2.10 and 2.11, the high within-agreement $\kappa = 0.8$ (table 2.8) and the good between-agreement show that listener three is the most representative listener. Another advantage of reducing the listeners to one representative listener is that it is far less time consuming to label a speech corpus.

2.4.2.1 Prominence assignment

The procedure for the prominence assignment was the same as for the training set. The text was presented on the monitor, and the spoken sentence was presented via headphones, with a maximum of three repetitions. The instruction was exactly the same as for the training set, apart from details about the number of sentences and the duration of the total experiment. The listener had to mark words in test sentences spoken with emphasis. Again, the listener could click on each word to mark it for prominence.

The result is that each word receives a binary labeling (0 = non-prominent, 1 = prominent). In total, listener three gives 3998 word prominence marks for these 1000 sentences, as specified in table 2.16 This results in, on average, four prominence marks per sentence. This is somewhat higher than for the sentences in her training set, which resulted in an average of 3.4 marks per sentence. This does not necessarily indicate poor consistency, but may rather show a small shift in threshold. Such shifts in threshold were also seen for the listeners who marked the training set.

Table 2.16: Absolute and relative numbers of words in the prominence judgments of the optimal listener three on the 1000 sentences of the independent test set.

Prominence marks	Freq.	%
Non-prom	6332	61.3
Prom	3998	38.7
Total	10330	100

2.5 Concluding remarks

We have presented an operational definition of prominence. Pilot experiments were carried out to find a useful operational definition for the remaining open questions. It was discovered that the prominence marks on words achieved by naive listeners are consistent enough and are useful for our further research. The 'open' instruction for the prominence marking task to the listeners (for instance not giving restrictions about how many words per sentence should be marked), show that this does not results in less agreement, but the differences within and between listeners can be dedicated to a threshold shift.
The cumulative marks per word can be used to express degrees of prominence, but the (arbitrary) scale from 0 to 10 can be reduced to a more useful 4-point scale by using the hierarchical clustering method. This leaves us with the following prominence marks in the training set:

- prominence degrees 0 to 10;
- prominence classes 0, I, II and III;
- for simplicity sake non-prominent (0 and I) and prominent (II and III).

In the test set we will frequently use a binary prominence distinction. However, we do not believe that prominence is binary or discrete, but is instead gradual. Only for simplicity's sake this gradual scale is reduced to a binary one. A similar approach is also found in Grover et al. (1997), Grover et al. (1998) and Buhmann et al. (2000).
In search for a binary division one might come up with the two classes 0 and 1-10 given the dendrogram in figure 2.1. However, token-wise the prominence class 0 is overrepresented, which partly explains its unique position in figure 2.1. We

therefore decided to combine 0 and I, as well as II and III to the binary non-prominent and prominent class, respectively (see table 2.14).

Only one representative and consistent listener marked the independent test set. Her binary marks thus result in:

- non-prominent (0) and prominent (1);

As we have chosen to mark word prominence, detailed information about the prominence distribution within a word is not available. Often the most prominent syllable coincides with the lexically stressed one, although this can differ and can shift to another syllable in the word. The choice for word prominence makes it easier for the listener to understand the task, because marks are given to meaningful elements (words). However detailed information about the differences in words as mentioned in the introduction 1.1 (*ACHTeruitgang (back / rear exit)* and *achterUITgang (decline)*) is not available. However, it was shown in Streefkerk et al. (1997) that it occurred hardly ever that a highly prominent syllable in a polysyllabic word does not correspond with the lexically stressed one.

In this chapter we described the operational definition of prominence. The listening experiment was designed in such a way that naive listeners mark words for prominence in sentences read-aloud. Generally, there are two sources of information the listeners use for this prominence perception: firstly the information which is transferred by the speech signal, and secondly their knowledge of the language.
In chapter 4 the acoustic correlates of prominence will be analyzed and discussed. In chapter 3 the textual correlates, which are related to the knowledge of the language (stored information) and prominence will be investigated and used for prominence prediction.

3

LEXICAL AND SYNTACTIC CORRELATES OF PROMINENCE[1]

Abstract

This chapter describes the relationship between perceived prominence and lexical / syntactic features, such as word class (Part-of-Speech), number of syllables, and the position of the word in the sentence. These relationships are formulated as heuristic rules, which predict the location and the degree of prominence in a sentence. The performance of the prominence / non-prominence predictor was found to be 81% correct when used on an independent test set. The ability to predict prominence with features that are derived from textual input can be applied in the fields of speech technology and particularly in speech synthesis.

[1] Parts of this chapter were published in Helsloot & Streefkerk (1998), in Streefkerk et al. (1999 a) and in Streefkerk et al. (2001).

3.1 Introduction

Speakers place prominence on particular words or syllables and this placement is guided by syntactic and lexical information and / or by semantics and pragmatics. The key issue here is to discover which of these syntactic and lexical features are the most effective guides. For example, when only syntax and the lexicon guide the placement of prominence, the result is a kind of neutrality in the style of the sentences spoken. We call this a default pronunciation of the sentence. Most commercial Text-to-Speech systems produce these more or less neutral sentences, since the context in which the sentence appears, is not known. But even without further context, sentences should preferably show some form of focusing.
"Fluency" (Fluent Dutch Text-to-Speech system Version 1.0, http://www.fluency.nl) a commercial Dutch Text-to-Speech system provides a 'default' realization of, for instance, the sentence *De nieuwste rage in Deventer is het schieten vanuit rijdende auto's (The latest craze in Deventer is shooting from moving cars)* as shown in figure 3.1. To make words prominent, the system produces a pitch movement on all

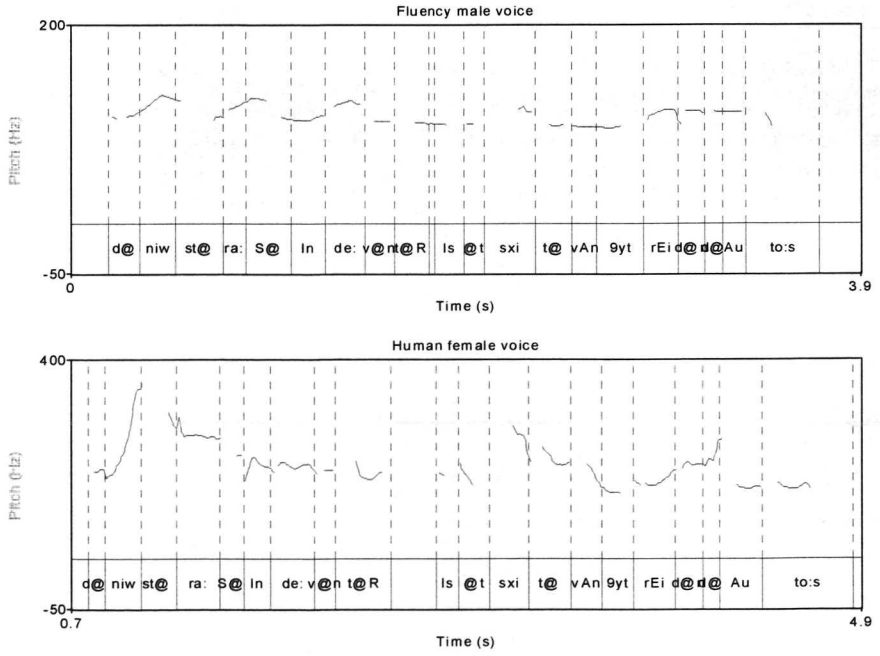


Figure 3.1: Pitch movements and segmentation of a *sentence De nieuwste rage in Deventer is het schieten vanuit rijdende auto's (The latest craze in Deventer is shooting from moving cars)*, spoken by the Text-to-Speech system Fluency (top) and spoken by a woman (bottom).

the content words, i.e. *nieuwste (latest)*, *rage (craze)*, *Deventer (city in the Netherlands)*, *schieten (shooting)*, *rijdende (moving)* and *auto's (cars)*. The system realized lexical stress by lengthening the duration of the vowel concerned, which also causes differences in prominence. However, in the same naturally spoken sentence from our Dutch Polyphone database (from a female speaker) only pitch movements are visible on *nieuwste (latest)*, *schieten (shooting)* and *rijdende (moving)* (see Figure 3.1). This illustrates the possibly large difference between naturally spoken sentences and those produced by a synthesizer. The naturally spoken sentence is enunciated with only one pitch movement on one of the three adjacent content words: *nieuwste rage (in) Deventer (latest craze (in) Deventer)*, whereas the synthesizer places a pitch movement on all three adjacent words. In Figure 3.1 it can also be observed that the durations of the syllables differ, which causes differences in prominence. These large differences are not only dedicated to things like speaker variability and gender. We believe that a better prediction of prosodic parameter on the basis of lexical and syntactic information is possible.

In principle, any word in a sentence can be prominent. Special meanings can be given to sentences in specific situations. Even Articles can be prominent, as in the following sentence: *Hij zei niet HET boek, maar DE boek (He didn't say the (neuter) book, but the (non-neuter) book)*. The Articles *de* and *het* are the most prominent words in this sentence. In normal conversation, however, those parts of speech that are most important for conversation are prominent. Articles are hardly ever prominent, but putting these words into contrastive / narrow focus as these two elements *de* and *het* stand in contrast to each other can highlight even those parts of speech.
In this study, however, the notion of focus is of secondary interest; primarily we want to point out that, in general, words containing new information and / or being important for the communication receive prominence and are less predictable from the context, whereas words with given information are more predictable. There is no need to highlight these words, as their message has already been announced.

From literature much is known about pitch accent and lexical stress. One stream of linguistic research deals with the relationship between lexical and syntactic information and the theoretic notions of stress and accent (for instance Chomsky & Halle, 1968). Another stream within linguistic research relates stress and accent with semantic and pragmatic information (for instance in Bolinger, 1972). Bolinger points out that accent placement is not so much a matter of syntax and word structure but a matter of information structure. He formulated the hypothesis that accents can only be predicted if the hearer is a 'mind-reader'. The exact placement of accent, and even more the exact placements of different prominence degrees in a sentence are determined by the speaker. It depends on how a speaker chooses to present information, and in this sense prominence is unpredictable as long as one is not a 'mind-reader'.
Prominence is a cluster of pitch accent and lexical stress. The computer is not a 'mind-reader', so the prediction of prominence under all conditions will be impossible, but most speakers still behave according to common linguistic

knowledge. Consequently, the prediction is possible up to a certain level Prediction of prominence could only be done for word types and not per word token. The aim of this chapter is to investigate to what extent we can predict prominence solely from textual information, coming as close as possible to a 'default' prominence prediction. The notions of pitch accent and lexical stress are of secondary interest.

3.1.1 Relevant studies

Two data-driven studies by Lea (1980) and Altenberg (1987), are discussed in section 3.1.1.1. These two studies deal with the relationship between stress and accent and word class based on empirical data. The more application-oriented studies are discussed in section 3.1.1.2.

3.1.1.1 Data driven studies

Lea (1980) describes the relationship between word categories and the amount of perceived reduction. He defines a continuum from stressed to reduced syllables and uses a perception experiment to define the amount of reduction. Five listeners were asked to judge if a syllable is stressed, unstressed or reduced. Adding up the judgments of all five listeners, a number on a scale from –5 (most reduced) to +5 (most stressed) is given to each syllable. The results show that various word categories can be arranged along a scale from most reduced (such as Articles, Conjunctions, Prepositions, Auxiliary Verbs, Pronouns), to most stressed (such as main Verbs, Adjectives, sentence Adverbs, Nouns, Quantifiers and command Verbs). It is worth pointing out that, generally speaking, most function words are perceived as reduced and most content words are perceived as stressed. This finding is used in early and in present-day Text-to-Speech synthesizers by simply giving all content words an accent. This approach tends to place an accent on too many words. Giving all content words a pitch accent results in a 'neutrally' spoken sentence. Although the sentences are comprehensible, such an abundant accent placement does not sound very natural.

In Altenberg's (1987) research various speech corpora (containing dialogues, radio speech, and face to face conversation) were labeled by hand for word classes and for prosodic information, expressed in prosodic labels ('zero', 'stress', 'booster' and 'nucleus') that indicate an increase of prominence ranging from 'zero' to 'nucleus'. Altenberg (1987) found that function words are only labeled 1-4% of the time with high prominence. Labels indicating high prominence were mainly found in the categories of content words. The study also shows that the various categories of content words can behave very differently when it comes to receiving prominence. Adverbs, for instance, were labeled as zero 23% of the time whereas Nouns were labeled zero only 8% of the time. Altenberg's study shows that, generally, the word class could be ranked for its ability to be accented. This means that there are words from a word class, which are hardly ever accented and other words from another specific class, which are almost always accented.
Such findings together with the 'hierarchy of ability to be accented' can easily be used and can actually be implemented in a synthesizer to optimize the accent

placement. The word class labeling must ideally be done automatically for this purpose, because for speech synthesis hand-marked word classes are not available either.

3.1.1.2 Application-oriented studies

We limit ourselves in this section to studies that deal with the prediction of stress and accent placement for Text-to-Speech systems, particularly for Dutch. An important requirement for such an application is that the textual information must be derived automatically, and the prediction should be easily implementable in a speech synthesizer.

Baart (1987) developed two algorithms to predict sentence accent for a Dutch Text-to-Speech system. The first, very simple, algorithm is based on the distinction between function words and content words. A list of unaccentable words is made; all words not occurring in this list are, accented. How this list was made is not described, nor does it cover the whole set of function words. The algorithm does not detect either the difference between homonymic words, such as the Noun *het weer (the weather)* and the Adverb *weer (once again)*. This algorithm can be used fully automatically and independently from a syntactic parser.
The second algorithm is based on syntax and focus and requires more information than simply the word class. Instead, the full phonological surface representation of a sentence and the domain of focus are used. This means that both the syntax and the lexical information are available. This additional information, such as the focus domain, had to be derived by hand and therefore this algorithm could not be used fully automatically.
Both algorithms were tested in a Text-to-Speech system. The accent labels were translated to specific acoustic parameters and listeners evaluated the resulting sentences. Results of the tests showed that the second algorithm is better as it uses more accurate and more specific linguistic information than the first. Research by Dirksen & Quené (1993) and Quené & Kager (1993), described in the next two paragraphs, is based on Baart's (1987) work and describes the design of a full system for synthesis applications.

Baart's (1987) first approach was analyzed in more detail in Quené & Kager (1993). The set of unaccentable function words was improved and extended with specific unaccentable content words (see for more details Quené & Kager, 1993). In principle all content words were accented, but some Verbs and other words which were considered as having less semantic information (such as *maand (month)* and *jaar (year)*) were excluded. As the algorithm still produced too many accent labels, the authors have devised two rules to remove some of them: the middle one of three adjacent content words is de-accented, and words which reliably convey given information are also de-accented, for instance epitheta before a proper name *(koningin Beatrix (queen Beatrix))*. Additionally, the authors developed special rules for the accentuation of Verbs. The algorithm gives a binary accent value: either plus or minus accent. Differences in degrees of prominence are not taken into account.

Dirksen & Quené (1993) designed a system, which is also basically founded on Baart's (1987) second algorithm. The accent assignment and phrasing is based on syntactic information presented in the form of a metrical tree augmented with focus markers. The information needed to predict the accent placement was not obtained automatically. However, some metrical aspects were also implemented, for example the system deaccented one of the two adjacent accented words. Results similar to the evaluation of Baart's second algorithm can be expected.

Hirschberg (1993) developed an algorithm for assigning pitch accent, which is used in the Bell Laboratories Text-to-Speech System. Several corpora were labeled for prosody according to TOBI (Pierrehumbert, 1980). Part-of-Speech, discourse information such as focus, and given and new information contribute to improve accent prediction. Hirschberg achieved a rate of 80-98% correct accent assignment for different speech corpora. A perceptual experiment to test the systems for the prediction of accent was not conducted, so no comment can be made concerning the increase of acceptance and/or the naturalness of the sentences with accent placement predicted by this sophisticated algorithm. Ultimately this algorithm provides only the prediction of an accent label, which then still has to be realized as a pitch accent in synthetic speech. This sometimes has consequences for the duration of the syllable. Hirschberg (1993) did not pay attention to the fact that words are perceived as having different degrees of prominence. Hirschberg & Rambow (2001) describe a more sophisticated approach. Adding tree-based syntactic dependency, e.g. the dependency between Auxiliary verbs and full Verbs or Determiners and Nouns, brings further improvements for boundary prediction, however, not (yet) for accent prediction.

Ross & Ostendorf (1996) used probabilistic approaches, which differ from the rule-based approach of Quené & Kager (1993). These probabilistic approaches show an increase in the correct prediction of prosodic labels. Moreover, they take into account different degrees of prominence. Ross & Ostendorf (1996) used computational models to predict accent location, symbolic tones such as H*L, and the relative prominence of these tones. The prediction of the relative prominence is limited to the F_0 peak height. As input Ross & Ostendorf used prosodic phrase structure information as to whether a word conveys given or new information, and Part-of-Speech labels plus information from the dictionary as input to accent assignment. All required information except boundary information is extracted automatically. In this study, Ross & Ostendorf argued that boundary location and pitch accent placement should ideally be predicted at the same time. But to simplify the problem they used hand-labeled boundaries to predict pitch accent. The exact location of the boundaries makes the prediction of accents easier. The pitch accent prediction was 82.5% correct using all available information. Results showed that four factors are important in the prediction of pitch accent: lexical stress, the number of syllables since the last pitch accent, the syllable position within the word, and the content word versus function word distinction. The pitch accent type and the number of pitch accents in the phrase are the most important cues for prominence.

Perceptual evaluation experiments were also performed. Listeners compared three different synthesized versions of one sentence. Version one contained their predicted accents, version two the hand-labeled pitch accents, and version three was the default realization of the synthesizer. In the experiment the listeners marked all three versions with 2.70 to 2.74 on a naturalness scale of 1 to 5, with 1 indicating the most natural. This result shows that synthetic speech is far from natural, and that even hand-labeled accents do not make sentences sound natural.
Bulyko & Ostendorf (1999) dealt with the prediction of gradient F_0 variation and its contribution to prominence. They used similar lexical / syntactic information as Ross & Ostendorf (1996) next to whether a word was new to the paragraph and / or whether a word was a part of a hand-marked named entity. A direct comparison of their results with previous ones is not possible since they are directly translated into acoustic features.

Three degrees of prominence (corresponding to accented or non-accented and to lexically stressed or unstressed) have been implemented in most present-day prosody generators. However, the implementation of different degrees of prominence in Text-to-Speech systems could improve the synthesized speech quality by making it sound more natural. This is the approach Portele & Heuft (1997) used. In their synthesizer they implemented different degrees of prominence. The algorithm was partly developed in perception experiments in which listeners had to mark on a 31-point scale the degree of prominence placed on syllables and the strength of the boundaries between words. By statistical analysis of a hand-labeled database, prosodic rules and a machine-learning algorithm to predict word prominence were developed (Widera et al., 1997). The information used to predict word prominence makes use of 21 different word classes (hand-labeled), the word class of the neighboring words, and their position in the utterance. The correct prediction rates of prominence are low: 41% correct at best.
The translation of the predicted prominence values to acoustic parameters was only performed for those syllables that have a mid / high prominence prediction (syllables with the value 17, on a scale from 0 to 30). Parameters concerning pitch movements were used to predict prominence for these syllables. Portele (1999) predicts height, position of the pitch movements relative to the vowel onset, and the steepness of the pitch movement. The differences in prominence were acoustically reflected in differences in F_0-peak height, but differences in duration and loudness were discarded. Some of the variability of 31 different prominence values was translated to acoustic parameters such as pitch, duration and loudness and was incorporated in a running synthesis system. The next step would be to translate not only the high prominence values to parameters concerning pitch movements, but to translate all the different prominence values including those concerning duration and loudness, to acoustic parameters.
Data-driven intonation modeling of both prominence on a 10-point scale and boundary strength on a 4-point scale is an approach (Buhmann et al., 2000; Fackrell et al., 1999) similar to that of Portele & Heuft (1997). In the research of Buhmann et al. (2000) and Fackrell et al. (1999) six European languages were involved. The initial labeling was only partly annotated by hand for prominence and boundary

strength; Vereecken et al. (1998) used automatic labeling for further annotation. They used both linguistic and acoustic features for automatic labeling. This combination brings further improvement for automatic prominence assignment. More than 110 acoustic features were used and about 70 linguistic input features, such as POS (Part-of-Speech), position of the word in the sentence, and position of the syllable in the word. Depending on the language, between 77% and 93% correct prominence classification is reached on a scale from 0 to 9 (with an accuracy of plus or minus 1 in comparison with hand labeled prominence). For Dutch the classification correct is 80%. However, no phonetic and linguistic knowledge about the relationship between those acoustic features and prominence is obtained with this large set of features.
For such large prosodically annotated databases, a probabilistic approach is possible. In this way, acoustic parameters such as F_0 contour and durations of speech segments can be predicted for Text-to-Speech systems. Automatic learning techniques, such as regression trees and neural networks, are fed with input features concerning orthography, phonetic transcription, segmentation, Part-of-Speech, word length, position of the word in the sentence, etc. Comparing the results of such approaches is difficult since it directly concerns the actually used acoustic parameters. Buhmann et al. (2000) found a correlation coefficient of 0.66 between the smoothed observed intonation contour and the predicted contour. Results for other acoustical parameters, such as duration and intensity are not presented in the above paper. A perceptual evaluation is reported in Fackrell et al. (1999). This evaluation shows that listeners prefer the automatically derived prosody model to the old hand-crafted prosody model of the Lernout & Hauspie Text-to-Speech-3000 synthesizer.

In summary, in the research discussed above, we see four different research lines: (1) algorithms are either based on linguistic knowledge, as in Baart (1987), or (2) on database research, as in Ross & Ostendorf (1996). Another possibility (3) is to develop algorithms by using perceptual labeling (prominence) as done by Portele & Heuft (1997) or (4) to predict the acoustical parameters of prominence and boundary strength directly as done by Buhmann et al. (2000) and Fackrell et al. (1999). They actually used both linguistic and acoustic features and do not have a perceptual evaluation.
The goal of our study is not only to show relationships between textual information and prominence judgments, but also to formulate these relationships in such a way that they can be expressed in heuristic rules to predict prominence. This rule set could then be used to enable more natural sounding speech synthesis. Therefore, all the desired textual information for the prediction of prominence must be derived automatically.

3.2 Pilot study to find lexical / syntactic correlates

In this section we describe a pilot experiment that was carried out to investigate the relationship between prominence and textual information. The pilot was based on a subset of 50 sentences. A main goal was to check if we could find a similar

relationship between prominence and word classes (POS, Part-of-Speech) as found in the research done by Lea (1980) and Altenberg (1987). In this pilot word class tags were assigned by hand for the following content word classes: Adjectives, Nouns, Verbs, Adverbs, additionally to the number of syllables and the position of the word in the sentence. This lexical / syntactic information was investigated with respect to prominence. A careful metrical and linguistic analysis of these sentences provided new insights, especially on (slightly) different word class definitions. The method used and the results obtained in this pilot experiment are described elaborately in Helsloot & Streefkerk (1998).

The pilot study confirmed that, in general, function words are perceived as less prominent than content words. Within these two categories we nevertheless found a considerable amount of variability. There were indications that the four content word classes are not enough to explain the variability, and that for instance Negations should be treated as a separate class. Our selection of the later used eleven word classes is based on the observations in the pilot study.
There also were indications that some of this within-category variability could be explained by the polysyllabic versus monosyllabic distinction. The question whether polysyllabic words are generally more prominent than monosyllabic words needs testing on a much larger set of sentences. It is necessary to investigate whether this effect is also valid within one word class, because this polysyllabic / monosyllabic distinction could perhaps also be ascribed to the fact that function words contain, on average, fewer syllables than content words.
In the pilot study the word class tags were assigned by hand. However, such tags will have to assigned automatically when predicting prominence for large sets of sentences. The discrepancy between hand-assigned word class labels and automatically derived labels could make the task of automatically predicting prominence more error prone.

3.3 Main experiment on lexical / syntactic correlates of prominence

So, in the main experiment presented below we will look more closely at the variety of word classes and we will try to explain some other aspects of the variability found. Therefore we will include combinations of word classes as well as the position of the word in the sentence.
In this analysis a much larger set of 1244 training sentences was used to analyze the relationship between perceived prominence and textual information (Part-of-Speech tags, number of syllables, and position of the word in the sentence). The results were used to derive a number of rules to predict prominence. These rules were evaluated with an independent test set of 1000 sentences. Ten listeners judged the 1244 read-aloud newspaper sentences of the training set, whereas only one representative listener judged the sentences of the test set (see for more details chapter 2). The cumulative marks of the ten listeners form degrees of prominence for the training set (11-point scale from 0 to 10). This scale was reduced to a binary scale (prominent versus non-prominent) since that was all we had available for the test set. The relationship between lexical / syntactic information on the one hand and prominence

on the other hand, had to be confirmed, further developed and elaborated into rules. First, we will describe the automatically derived word classes, then we will go into detail about the relationship between prominence and lexical / syntactic features, and at the end of this chapter the derived rules for prominence prediction will be described, tested and discussed.

3.3.1 Assigning lexical and syntactic features

The test set and the training set were automatically labeled for Part-of-Speech by a parser (Daelemans et al., 1996). The memory-based Part-of-Speech tagger is based on similarity reasoning. It compares a particular word in a particular context with the most similar case stored. Generally, two factors determine a Part-of-Speech tag: its lexical probability and its contextual probability. The lexical probability of a given word belonging to a given category is stored in the lexicon, whereas the contextual probability of a given word in a given context is stored in the case base. The lexicon and the case base were generated from a large corpus. Those lexical probabilities are only useful if a given word is known. Therefore unknown words are treated differently; the Part-of-Speech tag is guess-based on the form and / or context of the unknown word. It is reported that labeling fails in 5 to 10% of the cases (Daelemans et al., 1996). We used this memory-based Part-of-Speech tagger, because, although imperfect, to our knowledge it yields the best results for our goal. For our purposes nine word classes were distinguished (Noun, Adjective, Quantifier / Numeral, Verb, Article, Pronoun, Adverb, Preposition, Conjunction). Furthermore, based on the findings of the pilot study, the word class of Auxiliary Verbs is separated from the Verbs and forms a class on its own, just as Negations such as *niet (not)* do. In total we distinguish eleven different word classes. In the next section we describe the word classes used and give an indication of what kind of words belong to each word class.

Articles
The word class of Articles is relatively simple to define. Words as *de, het, der, des, een (the, the, of the, of the, a)* belong to this class.

Conjunctions
Conjunctions can be divided into coordinating Conjunctions such as *en, noch, alsmede, maar, of, want, dus (and, neither, as well as, but, if, because, therefore)* and subordinating Conjunctions such as *dat, sinds, toen, terwijl, zodat (that, since, then, while, so)*.

Prepositions
Normal Prepositions, such as *bij, tussen, voor (with, between, for)* belong to this class together with the group of prepositional Adverbs, which form a part of separable Verbs, such as *aangeven (to hand, to indicate)* and *voorkomen (to occur, to prevent)*.

Pronouns
Pronouns are divided into subgroups of relative Pronouns such as *die, dat* (*this, that*), personal Pronouns, such as *ik, wij, mij, ons, (I, we, me, our)*, possessive Pronouns *mijn, ons, jullie, zijn, haar (mine, ours, yours, his, her)* and the indefinite forms *iedereen, allen, allemaal, alles, iemand (everybody, all, all, everything, somebody)*. Based on the findings of the pilot study the pronominal Adverbs *er, daar* (*there)* were also included.

Auxiliary verbs
Only Auxiliary verbs of tense *zijn, hebben, zullen* (be, have, shall) and the passive voice *worden (be)* are put in this word class. Based on the observation in the pilot study the Auxiliary verbs of causality and modality are included in the full Verb class.

Verbs
This group contains all full Verbs plus the Auxiliary verbs of causality *doen, laten (do, let)* and modal Verbs *kunnen, moeten, mogen, willen (can, must, may, want)*.

Adverbs
Adverbs contain various subgroups. As already mentioned above, the prepositional Adverbs and the pronominal Adverbs have been shifted to the Prepositions or Pronouns, respectively. The Adverbs of place *waarheen, ginds (where, over there)*, time *toen, morgen, hoelang (then, tomorrow, how long)*, degree *nogal, graag (rather, gladly)*, and modality *misschien, wellicht (perhaps, possibly)* still belong to this word class.

Nouns
Proper names and nominally used Adjectives are put in the Noun category.

Numerals
The class of Numerals contains definite Numerals such as *beide, vier, driehonderd (both, four, three hundred*) and indefinite ones such as *veel, enkele, sommige, voldoende, minder (much, a few, some, sufficient, less)*.

Adjectives
The category of Adjectives contains the predicatively used Adjectives.

Negations
Based on observation in the pilot study Negations are put in a class of their own. It contains the words: *geen, niet, niets, nooit (none, not, nothing, never)*.

The speech material used was labeled automatically according to the word categories as described above.

3.3.1.1 Description and evaluation of the automatically derived Part-of-Speech (POS) tags

Since in this research the linguistic information used must become available automatically, the speech material was tagged automatically.
To obtain a clear idea of the number of errors made by the automatic parser, the training set (1244 sentences) was also tagged manually. The hand-corrected tags were initially obtained from CELEX. CELEX (Center for Lexical information, Celex@mpi.nl) provides all possible options for 'ambiguous' word tags. These hand-corrected tags were compared with the tags of the parser, which resulted in a number of 1057 out of a total of 13119 words that were incorrect classified. This is 8% of the total number (13119) of word tokens of the training set. Of these words 5970 (46%) are function words and 7149 (54%) are content words.
The total number of words distributed over the 11 word classes used is presented in table 3.1. The parser, described above, assigned the word classes. The categories of Articles, Prepositions, Verbs and Nouns are the largest, with for the training set 14.6%, 13.7%, 13.2% and 24.3%, respectively. The large number of Prepositions (13.7%) may be normal for this type of material, which was obtained from a newspaper. The smallest classes are the Negations (173, which is 1.3%), Numerals (327, which is 2.5%) and Conjunctions (434, which is 3.3%).

Table 3.1: Distribution of the eleven different word classes for the training set and for the test set, presented both in absolute numbers and in percentages. The word class was automatically assigned. The dotted line separates between function and content words (except for the Negations, which formally belong to the function words).

	Parser labeling training set			Parser labeling test set		
	Number		%	Number		%
Article	1912		14.6	1537		14.9
Conjunction	434		3.3	329		3.2
Preposition	1795	5970	13.7	1376	4645	13.3
Pronoun	1121		8.5	759		7.3
Auxiliary verbs	708		5.4	644		6.2
Verb	1734		13.2	1391		13.5
Numerals	327		2.5	268		2.6
Adverb	765	7149	5.8	558	5685	5.4
Adjective	977		7.4	805		7.8
Noun	3173		24.3	2521		24.4
Negation	173		1.3	142		1.4
Total	13119		100	10330		100

Table 3.1 also gives the parser labeling for the different word classes for the test set. The 1000 sentences used for test purposes contain 10330 words, which is on average 10 words per sentence (Std. Dev. 2.3). The number of content words is 5685 (55%) and of function words is 4645 (45%). These percentages per word class are more or less the same as for the training material.

Describing the relationship between these word classes and prominence is the next step, which will be presented in the following sections. Clear relationships are needed in order to be able to formulate heuristic rules to predict prominence.

3.3.2 Relationship between word class and prominence

Table 3.2 presents some general data on the training material in terms of means and median values of prominence marks (on a scale from 0 to 10), and their standard deviation. We also present the median values, because the distributions were askew at the edges of the prominence scale. The word classes were ordered by increasing prominence, indicated by the mean values, from Articles (mean prominence 0.1) to Adjectives (mean prominence 6.3). The standard deviation increases from 0.8 for Articles to 3.5 for Adverbs. In general, it is clear that the function words (Article, Auxiliary verbs, Prepositions, Conjunctions, and Pronouns) are the least prominent. Their mean prominence slightly increases from 0.1 to 1.5, but the standard deviation also increases. The median is still 0, thus indicating that these are not normal distributions, but we will discuss this later in this section.
The mean prominence values of the content words (Verbs, Adverbs, Nouns, Adjectives, Numerals) reflect that these words are more prominent than the function

Table 3.2: The mean, median and standard deviation of prominence per word class ordered by increasing prominence. The dotted line separates function words from content words, except for the Negations, which formally belong to the function words.

		Prominence		
Word class	Number	Mean	Median	Std. Dev.
Article	1912	0.1	0	0.8
Auxiliary verbs	708	0.3	0	1.2
Preposition	1795	0.4	0	1.4
Conjunction	434	0.4	0	1.2
Pronoun	1121	1.5	0	2.8
Verb	1734	2.6	1	3.1
Adverb	765	3.8	3	3.5
Noun	3173	5.6	6	2.8
Numeral	327	5.7	6	3.1
Negation	173	6.2	7	3.1
Adjective	977	6.3	7	2.8

words, which was already found in the pilot study (see Helsloot & Streefkerk, 1998) and by data in the literature (Altenberg, 1987 and Lea, 1980). However, we can be more specific, as the Verbs and the Adverbs have smaller mean prominence values (2.6 and 3.8) than the other content words. The words belonging to the Noun, Adjective, Numeral and Negation categories are perceived as prominent with mean values from 5.6 to 6.3 (see for more details table 3.2). A 'hierarchy of ability to be prominent' is already becoming apparent. The line in table 3.2 separates the function words from the content words. Negations do not strictly belong to the content words. As far as prominence is concerned, they behave similarly to content words.

As indicated earlier by the high standard deviations of prominence within each class (table 3.2), the variation within word classes is high and it is worth looking at the separate prominence distribution within each category.

These prominence distributions (on a scale from 0 to 10) per word class for function words are shown in figure 3.2 and for content words in figure 3.3. Function words are mostly those words, which are almost never marked as prominent (93% and 69% have zero prominence for Articles and Pronouns, respectively). The prominence histograms for Articles, Auxiliary Verbs, Prepositions and Conjunctions show that the higher the prominence marks the lower the percentages of the histograms (see figure 3.2). This holds also for Pronouns, but there is a small additional group that is marked with high prominence (seven, eight, nine and ten). Of the total number of Pronouns (1121, see table 3.2), 71 are marked with eight, nine and ten, which is 7% of the Pronouns. This can be explained by the fact that Subject-pronouns can be more prominent. Subject-pronouns are very difficult to detect automatically, which makes it difficult to use this information in a rule set for prominence prediction. The fact that only 71 words are involved, which is less than 0.5% of all the words from the training set, makes it less interesting too.

The prominence distributions with respect to content words look differently in several ways, as shown in figure 3.3. All content word classes show that the number of words never marked as prominent is less than 8%, except for the Verbs and the Adverbs. For the Verbs this number is 40% and for the Adverbs it is 27%. Verbs and Adverbs form a middle class, whereas the distributions of Nouns, Adjectives, Numerals and Negations are situated in the upper part of the prominence scale. Furthermore, Adverbs and, to a smaller extent, Numerals and Negations show bimodal distributions. 27% of the Adverbs are never marked as prominent, but the relative number of Adverbs marked by eight, nine or ten of the listeners is certainly high as well (22%). This once more confirms that the Adverb group is a difficult one, containing a variety of words that behave in different ways. Numerals and Negations show distributions with the maximum at the upper part of the prominence scale, but the number of Numerals and Negations judged by none or only one listener is still 7.6%, 5.3% and 7.5%, 7.5%, respectively (see also figure 3.3).

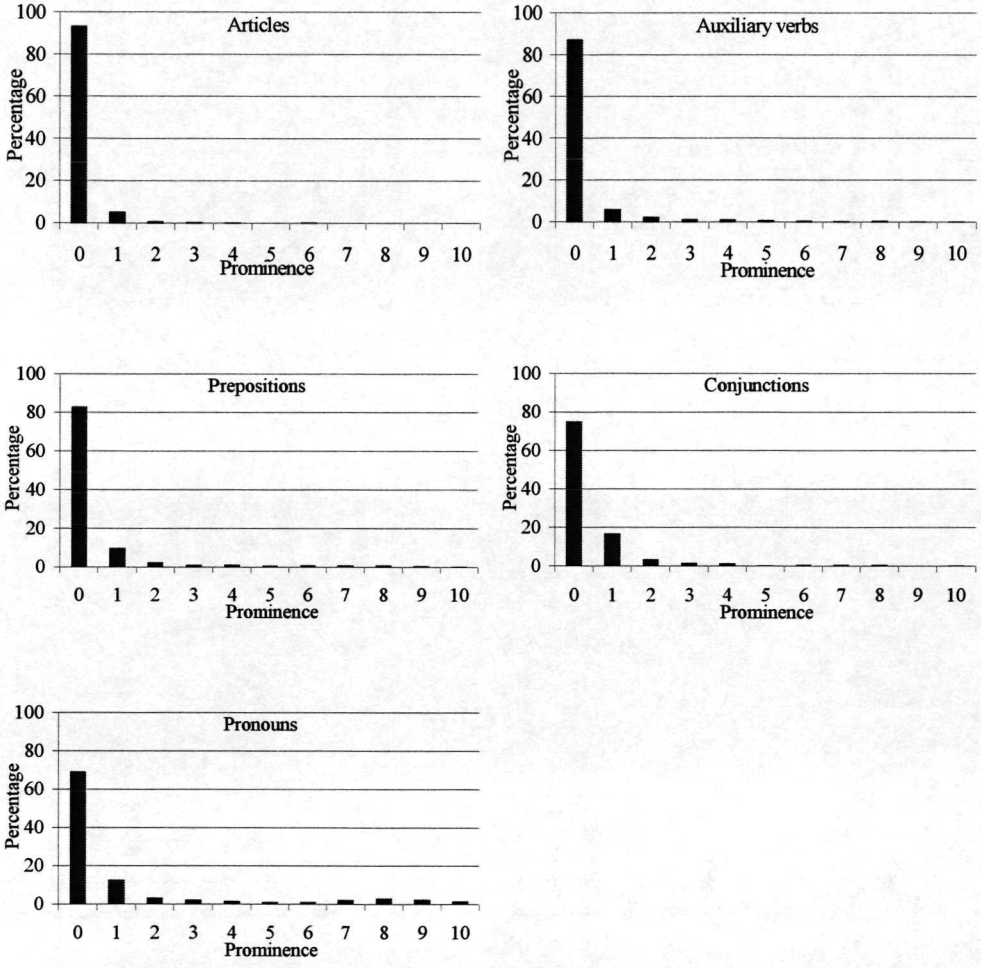


Figure 3.2: The distribution (in percentages) of the degrees of prominence (on a scale from 0 to 10) for the various types of function words.

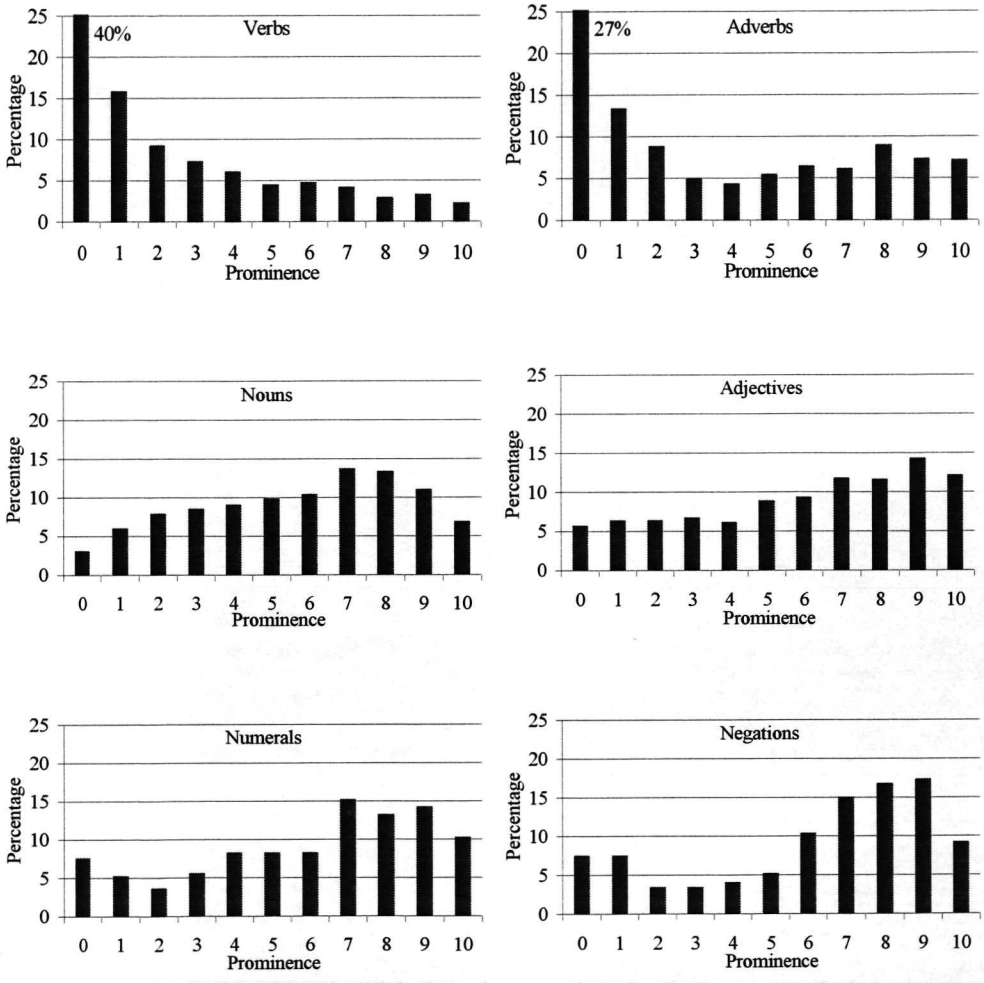


Figure 3.3: The distribution of degrees of prominence (on a scale from 0 to 10) for the various types of content words and Negations.

The prominence distributions of Negations and Numerals are also bimodal; in general these words are perceived as prominent but there are still a number of words that are never marked as such. An explanation for the non-prominent Numerals could be that Numerals referring to a current century are usually perceived as non-prominent. This occurs 14 times in our training material. Nine times these words are never or only once marked as prominent. As far as Negations are concerned, no explanation on a textual level could be found for the non-prominent ones, so the prediction of prominence for these special word groups is difficult.

3.3.3 Relationship between word length and prominence

An interesting relationship may be observed between the word length (number of syllables) and prominence, which cannot be exclusively attributed to the fact that more function words than content words are monosyllabic. Table 3.3 shows mean prominence value, median value and standard deviation of prominence, as a function of the number of syllables in a word. In conclusion, the longer the words, the higher the mean prominence values.

The difference between monosyllabic words (mean prominence value 1.4) and polysyllabic words (mean prominence values from 4.4 in the case of two syllables up to 6.4 in the case of six syllables) is considerable.

Table 3.4 presents prominence values of each word class split up for polysyllabic and monosyllabic words. Generally, it is true that the polysyllabic words tend to be more prominent than the monosyllabic words (see table 3.3). However, there is more variability within word classes. The difference between polysyllabic words and monosyllabic words is statistically significant ($p \leq 0.005$, tested with a student t-test for two samples) for Conjunctions, Prepositions, Pronouns, Verbs, Adverbs, and Nouns. Adjectives, Numerals and Auxiliary verbs are exceptions and there are no polysyllabic words at all in the class of Articles or Negations. The differences between polysyllabic and monosyllabic words are largest within Verbs, Adverbs, Conjunctions (only 15 polysyllabic words) and Pronouns.

Table 3.3: Mean prominence value, standard deviation and median value of prominence broken down by word length (number of syllables).

	Prominence			
Word length (num syll.)	Number	Mean	Median	Std. Dev.
1	7751	1.4	0	2.7
2	2769	4.4	4	3.4
3	1571	5.5	6	3.0
4	729	5.9	6	2.8
5	241	6.1	6	2.3
6	57	6.4	7	2.4
7	1	7.0	7	-

Table 3.4: Mean, median and standard deviation of perceived prominence of poly- versus mono-syllabic words per word class.

		Number	Mean	Median	Std Dev.
Article	Mono	1912	0.15	0	0.79
	Poly	-	-	-	-
Auxiliary verb	Mono	601	0.32	0	1.16
	Poly	107	0.46	0	1.25
Conjunction	Mono	419	0.32	0	0.86
	Poly	15	3.73	3	3.01
Preposition	Mono	1628	0.31	0	1.10
	Poly	167	1.64	0	2.60
Pronoun	Mono	947	0.86	0	2.12
	Poly	174	5.28	6	3.23
Verb	Mono	573	1.43	0	2.60
	Poly	1161	3.24	2	3.11
Adverb	Mono	496	2.83	1	3.26
	Poly	269	5.47	6	3.33
Noun	Mono	685	4.92	5	3.04
	Poly	2488	5.77	6	2.72
Numeral	Mono	152	5.70	6	2.98
	Poly	175	5.70	7	3.25
Adjective	Mono	165	6.30	7	2.94
	Poly	812	6.29	7	2.83
Negation	Mono	173	6.17	7	3.09
	Poly	-	-	-	-

The finding that polysyllabic words are generally perceived as more prominent may be connected to the 'metrical weight' of a word (Helsloot, 1995). The more syllables a word contains, the more weight this word carries in the utterance. Thus, the larger the metrical weight, the greater the possibility that the word will be prominent. This is especially true for Pronouns, Verbs, and Adverbs, which is shown in table 3.4 above. Figure 3.4 illustrates this even more clearly. We left the word class Conjunctions out of this figure because of the low number of occurrences of polysyllabic Conjunctions. The three word classes Pronouns, Verbs and Adverbs belong to the middle range in the 'hierarchy of prominence'. They belong to the group of words that are sometimes prominent and sometimes non-prominent. A distinction between polysyllabic and monosyllabic words may explain some of this behavior. Histograms for monosyllabic and polysyllabic words for these three groups are given in figure 3.4. Within a word class the percentages of a subclass (poly- or monosyllabic) add up to 100%. Most monosyllabic words lie in the lower range of the prominence scale, whereas the polysyllabic words are more evenly spread across the scale. The polysyllabic Pronouns and Adverbs move higher up the scale.

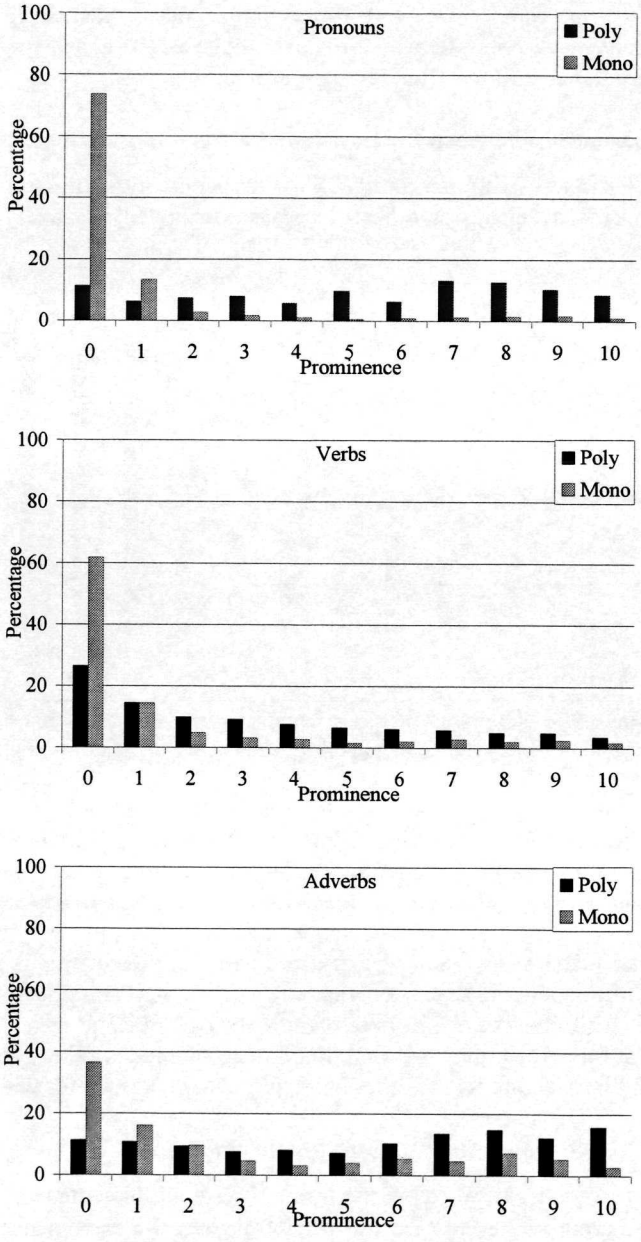


Figure 3.4: Prominence distributions in (percentages) of polysyllabic and monosyllabic Pronouns, Verbs and Adverbs. The absolute numbers are given in table 3.4.

In conclusion, monosyllabic Pronouns, Verbs and Adverbs carry more often low prominence in comparison with polysyllabic ones. Such differences between monosyllabic and polysyllabic words will be used in a heuristic rule system to predict prominence values from textual information.

3.3.4 Relationship between the position of a word in a sentence and prominence

Figure 3.5 shows a histogram in which the first content word in a sentence and other content words occurring at other places are displayed separately for prominence distribution.

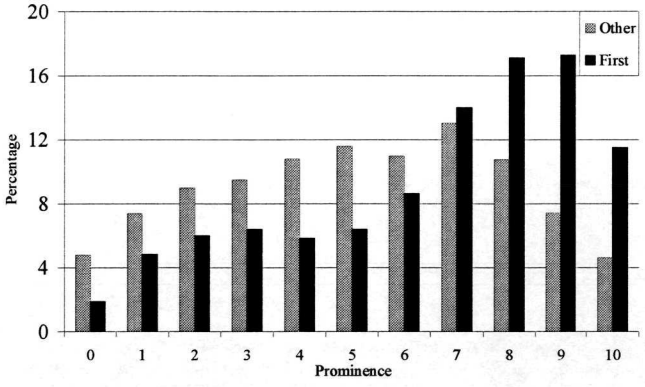


Figure 3.5: The prominence distributions of content words that occur at the beginning of a sentence as opposed to those occurring at other places in the sentence.

If words in the Noun, Adjective, Numeral, and Negation classes are placed at the beginning of a sentence, they tend to carry greater prominence than when placed at the middle and / or end of a sentence. The fact that these words are generally more prominent at the beginning of the sentence than at the end may be connected with the specific material we use. The first word carrying information is predetermined to be more prominent in sentences that are read aloud out of context, whereas in sentences with contextual information this may not be the case. This observation shows that rules for prominence prediction can be made optimal for a specific type of speech material, but on the other hand may also be unique for that material.

3.3.5 Adjective-Noun combinations and prominence

Of the total number of words in the training set of 1244 sentences, 24% are Nouns. This large group is spread over the whole range of the prominence scale, as shown in figure 3.3. There is a lot of variation within this word class. Some of this variation can be explained by taking a closer look at the combinations in which Nouns occur. The Adjective is much more prominent than the Noun in combinations where an Adjective is immediately followed by a Noun. Table 3.5 gives the mean, median and the standard deviation of prominence for the number of times Adjectives-Noun

Table 3.5: The number of occurrence as well as the mean, standard deviation and the median prominence value for Adjectives and Nouns found in Adjective-Noun combinations, and Adjectives and Nouns found in all other combinations.

	Number	Mean	Median	Std. Dev
Adjectives followed by a Noun	242	6.66	7	2.75
Nouns preceded by an Adjective	242	3.66	3	2.24
Adjectives in all other combinations	735	6.18	7	2.87
Nouns in all other combinations	2931	5.74	6	2.79

combinations, and Adjectives and Nouns occurring in all other combinations. The mean prominence values for Adjectives followed by a Noun (6.66) and in all other combinations (6.18) are more or less the same, as displayed in table 3.5. Nouns present a different story. The mean prominence values for Nouns when preceded by an Adjective (3.66), are much lower than the mean prominence values for the remaining Nouns (5.74), indicating that Nouns in such a combination are less prominent. The distributions of these two groups of Nouns are shown in figure 3.6. These two distributions do not correspond and confirm that the mean prominence values differ. These data show that the subgroup of 242 Nouns preceded by an Adjective could be predicted with greater accuracy.

It must be said that 242 Nouns is only about 8% of the total found in the training material, and about 2% of the total number of words. Whether or not this relationship can be formulated into a simple rule still needs to be tested and if so, also whether this rule improves prominence prediction.

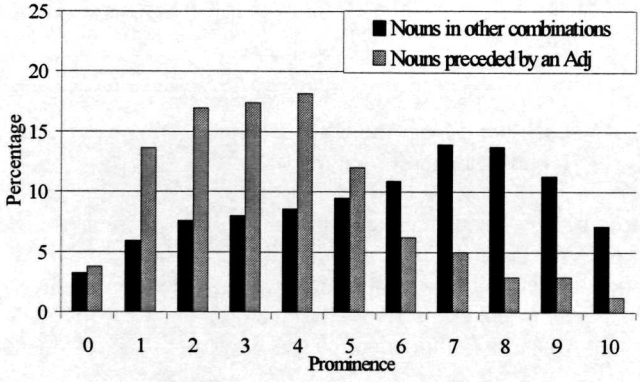


Figure 3.6: Distribution of Nouns, separated for Nouns which are immediately preceded by an Adjective and Nouns occurring in other combinations.

3.3.6 Algorithm with lexical / syntactic input for prominence prediction

In the preceding sections several relationships were presented between textual information and perceived prominence for a considerable total of 1244 sentences. There is a relationship between word classes and prominence. Function words are generally perceived as less prominent, whereas content words are perceived as very prominent. Words belonging to the word classes of Pronouns, Verbs and Adverbs are generally found in the middle of the prominence scale. We also discovered that the larger the number of syllables, the higher the mean prominence rating of that word. Furthermore, we saw that the first content word in a sentence was often perceived as very prominent. The Noun is generally less prominent in Adjective-Noun combinations than when found in other combinations. All this information has been used to devise an algorithm to predict prominence from text. This algorithm is based on a number of simple heuristic rules.

The following section describes how the relationship between lexical and syntactic information derived automatically from texts, and the prominence marks of the naive listeners will be combined into a set of heuristic rules for automatic prominence prediction.

To predict prominence by using the rules developed in this study, we devised a system similar to a 'metrical grid'. Units of words were put in a grid and then each word could be marked for prominence on different levels. The words can receive marks, which are indicated by an 'x', by applying various rules. After applying these rules, the grid is filled with marks (x) and consequently a prominence grid pattern emerges. Our rule system, described in more detail below, predicts up to four marks per word. The rules work additively, as will be clear from the example below.

Two heuristic rules can be formulated that reflect the general relationship between word class and prominence:

rule I : each content word receives one mark;
rule II : each word from the classes {Noun, Adjective, Numeral, Negation} receives an additional mark;

As an example of how these and subsequent rules are applied, we use one sentence from Polyphone: *Ik luisterde hoe de wind blies (I listened how the wind blew)*. We used a notation system similar to the metrical grid representation. In this example the words *luisterde* (Verb), *hoe* (Adverb), *wind* (Noun) and *blies* (Verb) receive a mark according to rule I, indicated by an x. Rule II gives the Noun *wind* an additional mark.

rule II					x	
rule I		x	x		x	x
	Ik	*luisterde*	*hoe*	*de*	*wind*	*blies*

The difference between polysyllabic and monosyllabic words should also be included in our set of rules. A simple suggestion would be that all polysyllabic words receive an additional prominence mark. Rule IIIa could then be formulated as follows:

rule IIIa: each polysyllabic word receives an additional mark;

Variations to this rule are possible. The difference between polysyllabic and monosyllabic words is largest for Pronouns, Verbs and Adverbs (for more details see table 3.4). A variant could thus be that only polysyllabic Pronouns, Verbs and Adverbs receive an additional mark. In this way, applying the rule would put polysyllabic Pronouns, Verbs and Adverbs at the same prominence level as Negations, Numerals, Adjectives, and Nouns, and they would receive an extra mark according to rule II. The prominence prediction of polysyllabic Adverbs and Pronouns with 5.47 and 5.28 mean perceived prominence, respectively, would then be the same as for Negations (6.17), Numerals (5.70), Adjectives (6.29), and Nouns (5.77). However, the actual mean prominence value for the largest group of polysyllabic Verbs is only 3.24. An additional mark is given to these Polysyllabic Verbs to avoid including these polysyllabic Verbs into one group with the other highly prominent words. Rule IIIb is thus reformulated in the following way:

rule IIIb: each polysyllabic word from the classes {Pronoun, Verb, Adverb} receives an additional mark, and each word from the classes {Noun, Adjective, Numeral, Negation} receives an additional mark;

A third variant is also possible. The difference between polysyllabic and monosyllabic words is statistically significant only for Conjunctions, Prepositions, Pronouns, Verbs, Adverbs, and Nouns. The difference is not statistically significant for Auxiliary Verbs, Adjectives and Numerals (see section 3.3.3). There are no polysyllabic words at all among our Negation class and Articles in our corpus. So the third variant of rule III would be that all polysyllabic words except Articles, Numerals, Negations and Adjectives receive an additional mark. All Adjectives and Negations must also receive an additional mark or else they would belong to the same predicted prominence level as the monosyllabic Nouns. This brings about the formulation of the third variant of rule III as follows:

rule IIIc: each polysyllabic word from the classes {Conjunction, Preposition, Pronoun, Verb, Adverb, Noun} receives an additional mark, and each word from the classes {Numeral, Adjective, Negation} receives an additional mark;

The next step is to determine which of the three variants most accurately predicts prominence. In our example the three variants differ in their prominence prediction for the monosyllabic Noun *wind*. The Noun 'wind' receives an extra mark only

when the IIIb rule is applied. The polysyllabic Verb *luisterde* receives an additional mark from all rules (IIIa, IIIb or IIIc).

rule III		x			(x)	
rule II					x	
rule I		x	x		x	x
	Ik	*luisterde*	*hoe*	*de*	*wind*	*blies*

Rule IVa is derived from our discovery that the first content word is more prominent than the other content words.

rule IVa: the first content word in the sentence receives an additional mark;

Variants of rule IV are also possible. One possibility would be to upgrade the first word predicted with two or three marks with one additional mark. Nouns, Numerals, Negation and Adjectives, plus polysyllabic Verbs and Adverbs belong to the group of two- or three-marks words.

rule IVb: the first word in a sentence with two or three marks receives one additional mark;

rule IV		x				
rule III		x			(x)	
rule II					x	
rule I		x	x		x	x
	Ik	*luisterde*	*hoe*	*de*	*wind*	*blies*
	0	8	0	0	7	2

The result is that the Verb *luisterde* receives an additional mark from both rules IVa and IVb and receives three marks in total.

The last rule we can formulate concerns the Adjective-Noun combination. As described in section 3.3.5, Nouns preceded by an Adjective are perceived as less prominent than Nouns found in other combinations. This then becomes rule number V for prominence prediction:

rule V: every Noun immediately preceded by an Adjective loses one mark;

Such an Adjective-Noun combination does not occur in our example sentence.

The perceived prominence values are given in the last row in our example sentence above. The two words *luisterde* and *wind* carry perceived prominence values of eight and seven, respectively. So in our example the words predicted with high prominence agree with highly perceived prominence. This is partly true for the less prominent word *blies*. This monosyllabic Verb *blies*, and the monosyllabic Adverb

hoe receive only one mark, but the prominence values are two and 0. As expected, the words with no marks at all are not perceived as prominent. All formulated rules are summarized below:

rule I: each content word receives one mark;
rule II: each word from the classes {Noun, Adjective, Numeral, Negation} receives an additional mark;
rule IIIa: each polysyllabic word receives an additional mark;
OR **rule IIIb:** each polysyllabic word from the classes {Pronoun, Verb, Adverb} receives an additional mark, and each word from the classes {Noun, Adjective, Numeral, Negation} receives an additional mark;
OR **rule IIIc:** each polysyllabic word from the classes {Conjunction, Preposition, Pronoun, Verb, Adverb, Noun} receives an additional mark, and each word from the classes {Numeral, Adjective, Negation} receives an additional mark;
rule IVa: the first word content in the sentence receives an additional mark;
OR **rule IVb:** the first word in a sentence with a level two or three mark receives an additional mark;
rule V: each Noun preceded by an Adjective is decreased by one mark;

The variants of rules III and IV are indicated with a, b, and c.
The application of these rules was partly demonstrated in an example in the previous sections and results in a prominence prediction system that predicts five levels of prominence on a scale from 0 to 4. The rules divided the words of each sentence into different groups. With the different variations of rules III and IV, six different combinations of rules can be put together. In order to get an optimal combination of rules, these different sets must be tested to find out which one is the most effective in predicting prominence. The different sets of rules are defined as follows:

Set A:	rule I	rule II	rule IIIa	rule IVa	rule V
Set B:	rule I	rule II	rule IIIb	rule IVa	rule V
Set C:	rule I	rule II	rule IIIc	rule IVa	rule V
Set D:	rule I	rule II	rule IIIa	rule IVb	rule V
Set E:	rule I	rule II	rule IIIb	rule IVb	rule V
Set F:	rule I	rule II	rule IIIc	rule IVb	rule V

Firstly, the various subsets that receive different numbers of predicted prominence marks must be made more clearly visible. It is difficult to see which set of words receives which number of marks, by applying the rules used in the various subsets. In Appendix table A 3.1, the subgroups of words receiving a given number of marks are presented. For instance, when applying rule set B or E, the function words receive no mark, but by applying set A or D only the monosyllabic function words receive no marks.
All rules described above were put into a simple algorithm in which the following textual input features are used:

-Part-of-Speech
-number of syllables in a word
-Adjective-Noun combinations
-position of the word in the sentence

This algorithm, in which the various rules are implemented automatically, assigned prominence marks to the training set. In order to effectively compare the results of each set of rules with the perceived prominence, a means of comparison must be found.
Basically, there are two problems. First, we want to compare different scales with each other. Secondly, sentences contain a different number of words; the greater the number of words in a sentence the higher the probability of error. Additionally, we want to compare the prominence contour of a sentence and not a comparison of individual words.
The predicted prominence values and the perceived prominence judgments were divided by their maximum scale value, being here 10 or 4, respectively. The result was that all values lie between 0 and 1. The total perceived prominence within a sentence was used for length normalization. The predicted prominence values were divided by this total. This allowed us to compare sentences of different word length with each other. Using this total (and not, for instance, the number of words per sentence) makes that falsely predicted prominence weighs more than falsely predicted non-prominence. The sum of the absolute differences per word is a value describing the goodness of fit. If predicted and perceived prominence fit perfectly, this value equals 0.

An example:

step I:							Maximum
Predicted number of marks (set B)	0	3	1	0	2	1	4
	Ik	*luisterde*	*hoe*	*de*	*wind*	*blies*	
Perceived prominence	0	8	0	0	7	2	10

Each prominence value, predicted and perceived, is divided by the maximum of its scale. All values then lay between 0 and 1.

step II:							Total
Predicted number of marks (set B)	0	0.75	0.25	0	0.50	0.25	
	Ik	*luisterde*	*hoe*	*de*	*wind*	*blies*	
Perceived prominence	0	0.8	0	0	0.7	0.2	1.70

Each value, predicted and perceived, is divided by 1.7, which is the sum of the prominence judgments, in order to fit the scales for the number of words and the number of prominence judgments and to allow comparison over sentences.

step III:						
Predicted number of marks (set B)	0	0.44	0.15	0	0.29	0.15
	Ik	*luisterde*	*hoe*	*de*	*wind*	*blies*

Perceived prominence	0	0.47	0	0	0.41	0.12

The absolute difference between the scaled predicted and perceived prominence gives an indication of the goodness of fit.

step IV:							Total error
	Ik	*luisterde*	*hoe*	*de*	*wind*	*blies*	
Difference	0	0.03	0.15	0	0.12	0.03	0.33

The total error of this example sentence is 0.33. Mean values of this goodness of fit could be calculated for each rule set by adding all values of each sentence and by dividing them by the total number of sentences.

The value of the worst fit depends on the total number of judgments given per sentence. This kind of normalization weighs correct prediction of highly prominent words and of non-prominent words in the same way.

The mean goodness of fit values of the prominence prediction over all training sentences are given in table 3.6. This table shows the effect of each rule incrementally: going from left to right an increasing number of rules were applied to obtain the prominence prediction. Only rules III and IV exist in more than one form and they vary depending on the set used. Consequently, rules I and II give the same match of perceived prominence and predicted marks for all sets. The results show that the more rules are applied, the better the fit is and this is true for all sets. Up to Rule III set B (0.821) and set E (0.821) are the best. Rule III b was applied in these sets. Rule III b only gives an extra mark to polysyllabic Pronouns, Verbs, and Adverbs. Up to Rule V set B is the best (0.616). Rule IVa (see rule distributions above) fits the perceived prominence judgments better and accordingly, set A (0.638), B (0.635) and C (0.639) are better than sets D (0.652), E (0.649), and F (0.652). This means that upgrading the first content word in a sentence results in a better fit than upgrading words that already have two or three marks. However, it must be mentioned that the different variants of Rule III and Rule IV do not bring a lot of improvement. Each additional rule improves the fit more than applying

Table 3.6: The mean normalized error of the different sets of rules applied to the training set. The rules are incremental, which means, for example, that values for + rule III are values for a prediction in which rule I, rule II and rule III were applied.

	Rule I	+ Rule II	+ Rule III (a, b, c)	+ Rule IV (a, b)	+ Rule V
Set A	1.130	0.873	0.829	0.638	0.619
Set B	1.130	0.873	0.821	0.635	0.616
Set C	1.130	0.873	0.827	0.639	0.619
Set D	1.130	0.873	0.829	0.652	0.628
Set E	1.130	0.873	0.821	0.649	0.625
Set F	1.130	0.873	0.827	0.652	0.629

Table 3.7: Absolute number of predicted prominence marks applied by the rules of set B crosstabulated against the prominence judgments of the listeners. The gray cells indicate the diagonal.

Perceived	Predicted prominence marks					
Prominence	0	1	2	3	4	Total
0		528	438	87	10	5977
1		195	311	135	13	1162
2	110		259	184	15	679
3	51		236	207	16	595
4	48	60		225	26	586
5	23	72		280	36	603
6	27	79	183		54	645
7	34	97	184		117	813
8	32	96	166	348		802
9	34	97	131	302		739
10	15	60	103	222		518
Total	5796	1480	2430	2673	740	13119

different variants of one Rule. Rule V improves the fit substantially, and the final result is that set B is the optimal solution to predict prominence on the basis of textual information only.

We will take a closer look at the results of the optimal set B. The resulting predicted prominence has been compared with the prominence judgments of the listeners and is presented in table 3.7, table 3.8 and in figure 3.7. Tables 3.7 and 3.8 present absolute numbers and relative numbers, respectively. The cells should be clustered around the main diagonal of the confusion matrix for an optimal prediction. The cells at the extremes of the prominence scale are more clustered around the diagonal than the middle prominence classes. The number 4914 given in the first cell in table 3.7 represents 82.2% of the total number of 5977 words never judged as prominent. These percentages are presented in table 3.8. More interesting is the second cell of the first row in table 3.7, which shows the number 528. The largest number of words with the prominence prediction of only one mark fall in this cell, but it is only 8.8% relative to the total number of words with a degree 0 prominence as presented in table 3.8. For the words with some degree of perceived prominence the quantities are more evenly distributed over the matrix. Only 118 of the 518 words that were perceived as highly prominent (marked as prominent by all ten listeners), were correctly predicted with four marks, which is 22.8%. The words with prominence degrees of 8 and 9 were predicted correctly (prominence mark 4) 160 times out of 802 (20.0%) and 175 times out of 739 (23.7%) respectively. Words with three predicted marks are distributed over the whole perceived prominence scale. The

Table 3.8: Matrix of relative numbers (percentages) of predicted prominence marks applied by set B with respect to the number of prominence judgments of listeners.

Perceived	Predicted prominence marks					
Prominence	0	1	2	3	4	Total
0	82.21	8.83	7.33	1.46	0.17	100
1	43.72	16.78	26.76	11.62	1.12	100
2	16.2	16.35	38.14	27.10	2.21	100
3	8.57	14.29	39.66	34.79	2.69	100
4	8.19	10.24	38.73	38.40	4.44	100
5	3.81	11.94	31.84	46.44	5.97	100
6	4.19	12.25	28.37	46.82	8.37	100
7	4.18	11.93	22.63	46.87	14.39	100
8	3.99	11.97	20.7	43.39	19.95	100
9	4.60	13.13	17.73	40.86	23.68	100
10	2.90	11.58	19.88	42.86	22.78	100

predicted prominence class with three marks covers perceived prominence classes from 11.6% for perceived prominence scale 1, up to 46.9% for perceived prominence scale 7. Fortunately, only 1.5% of the non-prominent words were marked with three marks. The cells for one and two predicted marks are more filled at the lower part of the prominence scale, and there is less spreading.

Figure 3.7 presents the relative numbers of predicted prominence over the perceived prominence degrees. Each column presents one perceived prominence degree and

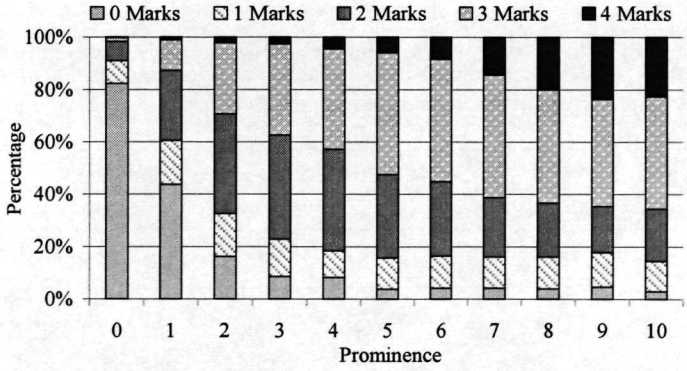


Figure 3.7: Predicted prominence values (set B) in comparison with the prominence judgments of listeners.

Table 3.9: Absolute number of occurrence, as well as mean, median and standard deviation of the perceived prominence per predicted mark.

Predicted prominence	Number	Mean	Median	Std. Dev.
0	5796	0.40	0	1.36
1	1480	3.15	2	3.40
2	2430	3.84	3	3.09
3	2673	5.85	6	2.77
4	740	7.50	8	2.25

the total of each column is 100%. The total number of words with perceived prominence degree can be looked up in table 3.7. For instance, 603 is the total number of perceived prominence at degree five. Figure 3.7 presents in principle the same data as presented in table 3.7. However, the data are presented as a diagram to illustrate more clearly that the extreme categories (no mark at all and four marks) are indeed found on the extremes of the perceived prominence scale, whereas words predicted with three marks are spread over the whole perceived scale.
Generally, the mean perceived prominence values agree with the predicted ones, but the standard deviation is still high. These values are given in table 3.9. The mean prominence value of the words marked with four marks is 7.5. These data reflect once more that the prediction of highly prominent words corresponds with the perception of the listeners. For the sets of words with one or two marks the mean prominence values are 3.15 and 3.84, respectively. For this set the standard deviation is the highest. The high standard deviation shows that more analysis will be needed to lower the large spreading. Maybe a more detailed look at the subgroups of Pronouns, Verbs and Adverbs, e.g. possessive Pronouns or relative Pronouns, will give indications of how to lower further the standard deviation. These values do not indicate that these sets should necessarily be treated as separate prominence categories.

Table 3.10: Absolute numbers of predicted prominence marks, relative to the total number of predicted marks within one class put in the prominence classes derived from the perceived prominence scale.

Perceived	Predicted prominence marks					
Prominence class	0	1	2	3	4	Total
0	4914	528	438	87	10	5977
I	618	306	570	319	28	1841
II	149	296	838	1014	132	2429
III	115	350	584	1253	570	2872
Total	5796	1480	2430	2673	740	13119

Table 3.11: The results of prominence prediction based on textual information for the training set. The predicted prominence marks and the perceived prominence classed have been agglomerated. The percentages of correct prediction are given in the last column.

Perceived	Predicted prominence marks			% correct
prominence class	0+1	2+3+4	Total	**82.0**
0+I	6366	1452	7818	81.4
II+III	910	4391	5301	82.8
Total	7276	5843	13119	

Using the clustered prominence categories rather than the original scores from 0 to 10 might be an advantage. This is discussed below.
The overall performance of predictive rules of set B for the training set of 1244 sentences is presented in table 3.10 and table 3.11 for two clustering methods. The original perceived prominence rate is chunked into four classes by means of hierarchical cluster analyses. (For details see, section 2.4.1.3.) Table 3.10 thus presents a 4x5-table, in which the non-prominence classification shows quite good performance: 4914 of the 5977 words are classified correctly. Prominent words cause more problems. Only 570 of the 2872 prominent words are classified with four marks, and the prediction in the middle prominence categories is even lower. However, by putting these data into a simple 2x2-table a classification rate of 82.0% correct can be achieved, see table 3.11. In this table it is also specified which categories are put together in the dichotomy between non-prominent and prominent.
As a conclusion we can say that the prediction of prominence is accurate for the extremes of the prominence scale, but that the middle section of the scale is much more difficult to predict. The following section deals with testing the prominence prediction rule on the independent test set.

3.4 Independent test of the prominence assignment rules

The algorithm for prominence prediction was tested on an independent test set of 1000 sentences. The same parser used for the training data automatically tagged POS labels for these sentences as well (see section 3.3.1 for more detail). With the

Table 3.12: Perceived prominence and predicted prominence marks for the independent test set.

Perceived	Predicted prominence marks					
prominence	0	1	2	3	4	Total
0	4001	841	930	516	44	6332
1	180	272	1284	1709	553	3998
Total	4181	1113	2214	2225	597	10330

Table 3.13: The predicted prominence marks and the perceived marks and an overall result in percentage correct.

Perceived prominence	Predicted prominence		Total	% correct
	0 +1	2+3+4	-	**81.2**
0	4842	1490	6332	76.5
1	452	3546	3998	88.7
Total	5294	5036	10330	-
Measure of agreement (κ)				0.62

help of the automatically derived word class labels, the number of syllables in each word, and the position of the word in the sentence, the various prominence levels were predicted according to optimal rule set B, as described in section 3.3.6. In order to test the generalizability of our set of rules, it is necessary to predict prominence on an independent test set. The procedure for matching predicted and perceived prominence for the test set shows differences with that for the training set. Only binary judgments of the 'optimal' listener are available for the test material. The prominence prediction on a 5-point scale and the prominence judgments of the optimal listener had to be matched, as presented in table 3.12 or as presented in 3.13. The middle section of the predicted prominence scale was distributed over prominence and non-prominence. Table 3.12 shows that prominence prediction with mark four is rare, but if it occurs then the word is almost always perceived as prominent as well. Similar to table 3.11, we also reduce for these independent test data the number of predicted categories from 5 to 2. A direct comparison with the binary perceptual scores then becomes possible. We observed to our satisfaction that the overall performance, even on this independent test set, can reach 81.2% correct classification. The exact data are given in table 3.13. As the data for the test set are derived from a single listener, it is possible to calculate the agreement between this optimal listener and the prediction achieved through the use of rules based on textual input. The resulting Cohen's Kappa of 0.62 is, in fact, better than the between-listener agreement, with a mean Kappa of $\kappa = 0.50$ (St. Dev. = 0.16).

3.5 Discussion and conclusion

In the previous sections, we have shown that it is possible to predict perceived prominence with an accuracy of 81.2 % on the basis of textual information only. The advantage of this kind of prediction is that most of the required information can be derived automatically. The prominence prediction is used on the following text features:

-Part-of-Speech
-number of syllables in a word
-Adjective-Noun combinations
-position of the word in the sentence

The agreement between the automatic prominence prediction and the optimal listener may not seem high ($\kappa = 0.62$), but is higher than the agreement between listeners (mean $\kappa = 0.50$). The performance on the training set is 82.0% correct classification (binary: in terms of prominent / non-prominent). With the present analysis of the data, appropriate rules have been formulated. These rules are simple and intuitively appropriate. A disadvantage is that some rules may be specific for the particular types of sentences used in this experiment. It could be the case that for a different type of speech material a different set of rules must be formulated.

In accordance with Fackrell et al. (2000), there is much prosodic variation between different text types. It is, for instance, very well possible that rule IV (every content word at the beginning of a sentence receives an additional prominence mark) is specific for this type of speech material. In the material we used, all sentences were read aloud without context, so it is possible that most speakers felt the need to highlight the first word carrying some kind of information. Given a text type one may be able to formulate text-specific rules and consequently to improve prominence prediction.

4

ACOUSTIC CORRELATES OF PROMINENCE[1]

Abstract

This chapter discusses the acoustic correlates of prominence, how to measure them, and how to extract them automatically from the speech signal. The chapter is structured in the following way: first a general description of acoustic correlates is given, followed by a detailed description of the relevant literature concerning automatic feature extraction. Then, an acoustical analysis of the speech material used in this study is given. The analysis mainly concerns the 1244 sentences that have already been marked for prominence (see chapter 2). Special attention is paid to the automation process of feature extraction.

[1] Parts of this chapter were published in Streefkerk et al. (1998), Streefkerk et al. (1999 a), Streefkerk et al. (1999 b) and Streefkerk et al. (2001).

4.1 Introduction

Information available about (the degree of) prominence can be used in several ways. For example, automatic prominence labeling may be useful for speech recognition applications, as it may provide additional information to the recognition process. The degree of prominence of a word may help to identify if a word is important for the communication processes or not. Additionally, prominent words can be used as islands of reliability. The tool, if fed with relevant acoustical information, assesses the degree of prominence of each word in the sentence. In principle, each word in a sentence can carry prominence, so for such an application the a priori probability as found in the training material is the basis of the assessment.

For instance, in dialogues, a useful application of prominence is to distinguish between sentences that contain the same words, of which the meaning changes with a shift in the position of prominence from one word to another: "*voor INstappen" (for getting on)* versus "*VOOR instappen" (get on at front)*, or an English example "in CAPable hands" versus "INcapable hands". For disambiguation it is important to know which of the two relevant words (or syllables) is the most prominent one. Out of context those words (or syllables) have statistically the same chance of being the most prominent one.

Having these two applications in mind, this chapter concentrates on acoustic features that indicate the nature of prominence (as judged by naive listeners). In other words, the main problem in this chapter is how to extract relevant acoustical cues from the speech signal and to study their distribution over prominence classes. In chapter 5 a selection of these acoustic features will be used as input to a neural network for prominence classification.

4.1.1 General description of possible acoustic correlates

Prominence marking, as defined in chapter 2, is a perceptual labeling process performed by humans. This process establishes correlates by using the speech signal (acoustic correlates) and knowledge of the language (here operationalized as textual correlates). The prominence correlates related to text were described in chapter 3, whereas the main topic in this chapter concerns the acoustical cues of prominence. In finding acoustic correlates we are faced with three main problems.
Prominence may be manifest in terms of a speech segment being louder, longer, more clearly pronounced, and containing pitch level changes. These are all perceptual / articulatory terms and are therefore difficult to measure from the speech signal. However, it is possible to measure physical characteristics in the speech signal. Following this line the first problem is to associate these perceptual / articulatory terms with acoustical measurements.
The second problem concerns the fact that prominence is always relative to its environment, and that measurements must take into account this relativity.
A third problem for this type of research concerns the unit (word, syllable, phoneme) of speech to be measured.

Numerous studies give evidence for several acoustic correlates of prominence. Generally, the following acoustic features can be associated with prominence:

- F_0 (Hz)
- duration (s)
- intensity (dB)
- spectral characteristics

Various remarks concerning these acoustical measurements can be made. In general, change in F_0, which is perceived as rising or falling pitch movements, cause words to be perceived as being more prominent than other words in which smaller or no changes in F_0 are measured. In several studies it has been confirmed that, for instance for English and Dutch, changes in F_0 are the most important tool in marking words for prominence (Ladd, 1996; 't Hart et al., 1990; Lehiste, 1970; Bolinger, 1958). Changes in F_0 are not absolutely related to the phenomena of prominence; they depend on the type of movement (rising and falling pitch movements in terms of the IPO intonation grammar (Hermes & Rump, 1994)) and must be interpreted relatively within one utterance (Terken, 1996; Hermes, 1995; Hermes, 1991). Adjacent pitch movements can also influence the prominence perception (Ladd et al., 1994). F_0 changes must be seen as relative to several factors: to the changes found in neighboring words, to the beginning or the end of a sentence ('t Hart et al., 1990), and to the speaker (Gussenhoven & Rietveld, 1998; Kraayeveld et al., 1991), and his / her emotional status (e.g., angry speakers general produce larger changes in F_0 than speakers who are bored (Mozziconacci, 1998)).
All these factors influence F_0. In our study we do not only have to cope with the factors mentioned above, but also with the restriction that the labeling / classification process, including the feature extraction itself, must be done automatically. Firstly, the F_0 measurement itself often introduces octave errors. Secondly, for our research we can only correct for factors that are automatically available, for instance, information about the vowel identity is available for our type of speech material, but the emotional status of the speaker is not.
The research described above concentrates on the shape of the F_0-curve, and not on the feature extraction in the individual syllable or word. Striking changes in the F_0-curve have been assigned to words (or syllables) later. Such research highly depends on a good F_0 measurement algorithm, whereas a proper alignment with the words (syllables) is also a complicating task. In our research we want to concentrate on the more local acoustic features, which can be found in the individual syllables or words. Our approach will be to find in the appropriate syllable (or word) features that will provide information of the prominence of that syllable (word) and, furthermore, will be suitable for an automatic classification task.
According to the literature the duration of the various speech units (words, syllables, phonemes) is also related to prominence. Generally, a long syllable or vowel duration corresponds with increased prominence, but this duration also depends on several other factors: the speaking rate, lexical stress, intrinsic segment properties, the number of segments that constitute a syllable and / or a word, finality (Cambier-Langeveld, 2000), and the following segments. In the acoustical analyses of this

chapter it must be investigated for which of these factors normalizations are helpful in order to have proper durational features of prominence.
The intensity of the vowels and / or syllables is reported in the literature as being important also for prominence, but there are also several other factors, which influence the intensity of vowels and syllables. First, there is lexical stress, which is basically a property of the lexicon, and influences the intensity of the whole syllable and especially the intensity of the vowel (van Kuijk & Boves, 1999; Sluijter & van Heuven, 1996). Second, there is the intrinsic intensity of vowels: open vowels are more intense than closed vowels (Lehiste & Peterson, 1959). A third point that has to be taken into account concerns sentence finality: the intensity decreases towards the end of a sentence.
The last notable acoustic correlate concerns the spectral characteristics of the vowel and / or consonants. The phenomenon of spectral reduction of vowels (F1 / F2 values shift to the middle of the cardinal vowel diagram) related to prominence is discussed in van Bergem (1993). Spectral reduction of consonants is related to the spectral center of gravity (the 'mean' frequency in semi-tones, weighted by spectral power), see van Son & Pols (1997).
The possible acoustic correlates of spectral characteristics will not be dealt with in this study, partly because the quality of recording over the telephone is probably not good enough to permit the measurement of reliable spectral characteristics, and partly because the influence of other features is even bigger here.

4.1.2 Relevant studies on automatic feature extraction

Several studies discuss the prosodic notions of accentuation and phrasing and the extraction of prosodic features. We here focus on the studies concerning automatic prosodic feature extraction of accentuation and (lexical) stress in order to detect prosodic properties related to the prominence of words and syllables. The selected studies label accentuation differently: according to the IPO intonation grammar, or according to the TOBI-label system (Taylor et al., 1998; Wright & Taylor, 1997; Kießling, 1996; ten Bosch, 1993; Vaissière, 1989). Other studies label their material for focus (Petzhold, 1999; Heldner et al., 1999), or use the notion of prominence, (e.g. Wightman & Ostendorf (1994) use prominence marks at the syllable level).
Related research concerns the automatic detection of (lexical) stress. In van Kuijk & Boves (1999) an attempt was made to improve HMM speech recognition by lexical stress recognition from acoustic features. Hand-labeled stress marks (unstressed, stressed) are used in the research of Silipo & Greenberg (1999).
All these studies deal with automatic feature extraction for the detection of their prosodic features. In the next paragraphs we will briefly discuss various approaches to feature extraction. The results are not always comparable because of the differences in the speech material used and differences in initial labeling.
Studies concerned with the automatic detection and classification of pitch accents mostly concentrate on the shape of the F_0-curve (e.g., Taylor et al., 1998; Maghbouleh, 1998; ten Bosch, 1993). The most striking rises and falls are the basis for defining various features taken from the F_0-curve and their timing is generally related to vowel onsets. No further lexical information is used. The correct

recognition rate for a pitch / non-pitch accent decision for the two studies by ten Bosch (1993) and Taylor et al. (1998) is 81% and 74%, respectively. Vaissière (1989), who also tried to classify the different types of pitch accent for different speakers, already mentions that the surface realization of pitch accents differs widely among speakers. Frid (2001) investigated and predicted intonation patterns in terms H* L*. Via pitch contour stylization patterns of accented words were compared and put into 30 classes, which makes clear that a lot of variation in the intonation contour is possible. As indicated before, we thought that the concentration on the shape of the F_0-curve is not a useful way to detect prominence automatically. From the whimsical changes in the F_0-curve it is difficult to provide information about the prominence degree of each individual word or even syllable. As said before we want to concentrate on the features that are directly related to, and directly extractable from, the word (or syllable) concerned.

Kießling (1996), who aims at the recognition of pitch accents, uses features describing the F_0-curve as well as a large set of features concerning duration and intensity, and all of these also with respect to the previous and following syllables. He applied an effective method of making the absolute features relative to their direct environment. Even lexical information, such as lexical stress or the identity of the segments, or even the number of syllables in a word, was used. A total of 276 features per syllable were used for the classification task. The result for pitch / non-pitch accent decision with this large feature set is rather good, namely 83% correct, but, in our view, such a set of features is too large to analyze the specific contribution of the separate features. In Batliner et al. (1999) this large number of features is reduced to six, and they conclude that F_0 features are not more important than intensity or duration features. These findings convince us that pitch accent is not only achieved by accent lending pitch movements, but also by other acoustic correlates such as duration and intensity. Batliner et al. (1999) used additional information concerning the content-word / function-word distinction, lexical stress features, as well as the number of syllables in a word.
The studies by Batliner et al. (1999) and Kießling (1996) show how difficult it is to make clear what type of information (lexical / syntactic information or acoustical information) is used for classifying pitch accent. It is certainly clear that the duration of a word highly correlates with the number of syllables, and that the number of syllables correlates with the content-word / function-word distinction. As function words are less prominent than content words (see chapter 3) the type of information is not clearly defined.
Wightman & Ostendorf (1994) discuss the recognition of initially hand-labeled syllable prominence. They also use the syllable as a basic unit and extract different features concerning F_0, such as the mean F_0 in a given syllable, the mean intensity of a syllable, as well as features concerning the duration. Even pause duration is used. These features are also used relative to the syllable before and after the current one. Flags for lexical stress and finality of words were added to the feature set. They obtain a recognition rate of 83% for the prominent / non-prominent distinction.

The approach of Batliner et al. (1999), Kießling (1996) and Wightman & Ostendorf (1994) is interesting for our type of analysis and classification because they give several features describing the nature of pitch accent / prominence in the word (or syllable) concerned. Corrections for other influential factors appear to be possible, as well as a direct linkage to lexical information. In our research we will not take the properties of the surrounding syllables into account, because an interpretation of this relativity is difficult. These relative features prove to be very useful for a speech-engineering approach of prominence classification. For our research high recognition rates are not the only goal; we concentrate on the need to provide phonetic insight about which features are important for prominence classification. Which are the most useful features for our prominence detection and classification of prominence will be analyzed in more detail in the following part of this chapter.

There are two other studies worth giving more detail: the study of Silipo & Greenberg (1999) and that of van Kuijk & Boves (1999). The former authors classified initially hand-labeled stress (primary, secondary, unstressed) with the help of vowel measurements in a given syllable. They reached a recognition rate of 80% for stressed syllables and 77% for unstressed syllables (the recognition rate of the secondary stressed syllable is not reported), by using energy, mean F_0 value and duration of the vowel. Van Kuijk & Boves (1999) described interesting acoustic correlates for lexical stress (automatically annotated by means of a lexicon). They tested different features, especially those concerning the duration and the energy, such as the total energy of a vowel and the spectral tilt, and they used various normalizations. A correct classification rate of 72% is reported. Interestingly, the total energy seems to show high discriminability, but this is due to a combination of two features as the total energy is directly related to the duration of a given sound. An additional point is that most features show statistical differences for the stressed / unstressed distinction and seem to be dependent of vowel type.

In the speech material that we use in our own research there are several complicating factors, which make the measurements more difficult than in some studies described above:

- speaker variety
- different environments
- data recorded via the telephone

A further complicating factor is the fact that both the 'basic' acoustical measurement (F_0, duration and intensity) and the subsequent acoustic feature extraction must be done automatically.

The goal of this automatic extraction of acoustic correlates of prominence is not only to analyze which features are important for prominence and what are the relations of the acoustic features is to each other (mainly the topic of this chapter), but ultimately also to recognize prominent and non-prominent words automatically (as described in chapter 5).

4.2 Feature extraction

4.2.1 Segmentation and labeling of the speech material

The used speech material (Dutch Polyphone Corpus) has to be segmented into words, syllables and phonemes, as otherwise duration is not measurable, whereas F_0 and intensity otherwise could not be aligned to lexical units containing lexical information, such as vowel type or lexical stress. Of course it is known that segmentation by hand is a very time-consuming procedure and that even hand transcription shows certain systematic errors (Cucchiarini et al. 2001; Eisen & Tillmann, 1992). However, we aim, as much as possible, to have automatic procedures. Automatic segmentation facilitates the processing of large amounts of speech material, but at the same time it introduces several types of errors. Some of these errors are discussed later in this section, but first we will turn to the general procedure for automatic segmentation of our speech material.

All the phonetically rich sentences from the Dutch Polyphone Corpus have been orthographically transcribed, which means that it is known what the textual content on word level of each given speech signal is. Background noise, non-speech sounds, and noise from the speaker himself / herself (lip smacks) are labeled between brackets. The example below shows one of the sentences recorded.

1) *Hij heeft twee argumenten voor zijn stelling. (He has two argument for his proposition)*

2) [mouth_noise] *hij heeft twee argumenten voor zijn stelling* [bg_noise]

In the subsequent section this specific sentence serves as an example sentence. If a speaker stutters then these hesitations are also described. Additional information about hesitations, speaker and background-noises has been added by hand. The enriched transcription will facilitate automatic segmentation. We realize, of course, that in real speech technology applications, such a complete transcription is generally not available. However, we used this kind of information in order not to unduly complicate the segmentation problem.

The transcription of the sentence plus background noise and the noise from the speaker is available for the speaker.
The phonetic transcription was added with the help of the standard pronunciation lexicon of the Dutch language (CELEX). For each sentence the standard pronunciation was looked up, resulting in a chain of phonemes, syllables and words as presented in 3) below, which represents the normative phonetic transcription of this sentence in SAMPA notation (see Appendix A 4.1).

3) [mouth_noise] HEi he:ft twe: AR-Gy-mEn-t@ vo:R zEin stE-lIN[bg_noise]

We used an HMM speech recognizer to align the speech signal to the concatenated sequence of phonemes. In this procedure there are, however, various sources of error. First, the system supposes that the pronunciation will be the standard one, which is improbable because of differences in speaking style, gender, and in the regional background of the speakers. Second, although the alignment allows an optional silence to be inserted between subsequent words; it is not possible to insert or delete phonemes. In other recognizers alternative pronunciation possibilities are sometimes implemented together with the possibility of making deletions and insertions. However, such extensions may introduce complications, especially for training the HMM-model, so we therefore opted for a simple variant.
In our example sentence the following problems occur: the two /t/'s in the two connected words *heeft twee (has two)* are most probably reduced to one /t/. Within words the pronunciation lexicon may solve such de-geminations. For instance, in the word *aannemen (take, accept)* the two /n/'s are pronounced as one /n/, and CELEX correctly transcribes this word as /a:-ne:-m@/. However, such assimilation and co-articulation phenomena are not supported beyond word boundaries. Another complication is the deletion of the /n/ at words ending with /@n/. This is a debatable point. For the word *argumenten (propositions)* in our example 3), the lexicon (CELEX) gives an obligatory n-deletion at the end of this word. Under certain conditions and in certain regions of the country this /n/ deletion at the end of a word may be correct, but there are regions in the Netherlands where this kind of deletion is not common.
Other problems concern reduction, speaking rate, style etc. An example of reduction is found with the possessive pronoun *zijn (his)*. In example 3) this is transcribed as /zEin/, however, it is often reduced to /z@n/.
There are several other problems one could think of, such as for instance in *er was eens ... (once upon a time)*, which forms the beginning of fairytales, and which could be reduced to /wAz@z@n/ or even /wAz@s/ or /wAs@s/. The frequent insertion of a schwa in the word *film (film)*, which is then pronounced as /fIl@m/ is an example of an inserted phone, which is not represented in the canonical pronunciation either.
We will next discuss the training of the HMM recognizer and we will compare the automatic segmentation of the example sentence with the hand segmentation.

4.2.1.1 Training of the HMM-recognizer

About 4500 sentences (a subset of 3 CD-ROMs with the sentences of a total of 900 speakers) were used to independently train HTK an HMM recognizer[2]. The standard pronunciation for these 4500 sentences was taken from CELEX. This resulted in a chain of phonemes per sentence, which were used for training. The phone models of this recognizer were used to localize the segment boundaries or, in other words, to align the given string of phonemes. 38 different symbols (16 vowels and 22 consonants) were used for speech, mostly according to the SAMPA notation (see Appendix A 4.1). This list of Dutch phonemes has been enlarged to include 5 symbols for noises and silence.

[2] with the much appreciated help of Xue Wang (Wang, 1997)

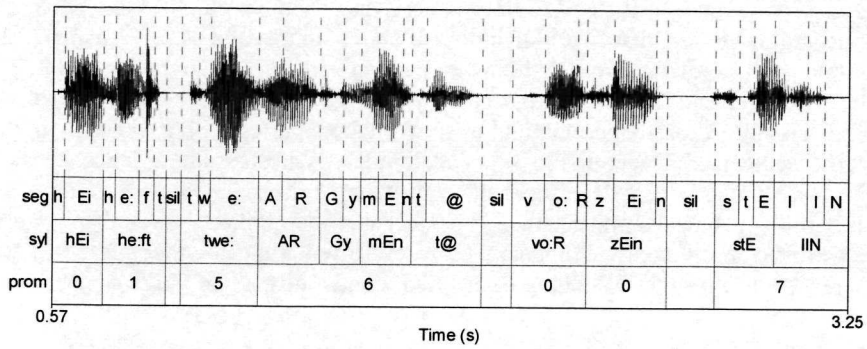


Figure 4.1: Oscillogram of the sentence *Hij heeft twee argumenten voor zijn stelling (He has two arguments for his proposition)* with accompanying automatic phoneme segmentation and the related syllable and word segmentation tiers. The perceptual prominence scores on the word level are indicated in the lowest tier.

In total 58 separate phoneme models were trained: 5 for non-speech, 16 models for vowels, 15 additional models for vowels in stressed position (the schwa cannot occur in a stressed position), and 22 models for consonants. In order to distinguish these models, lexical stress is defined using CELEX, with an added restriction that the function words as defined by van Wijk van & Kempen (1979)[3] never bear lexical stress. With the analyses in chapter 3 the definition of function words could be transformed into a more suitable one, as in principle only the segment boundaries are definite.

Each phone model contains 5 states of which states 2, 3, and 4 contain a self-loop and can be skipped. Each HMM state corresponds with a mixture of 8 Gaussian PDF's (probability density function). The HMM models use 12 FFT-based MFCC (mel frequency cepstrum coefficients) and the log energy as input vectors, which results in 13 parameters per frame. The delta and delta-delta parameters are also used. For each frame a hamming window of 25 ms is used and a frame shift of 10 ms. The minimum duration of each segment is 30 ms, as the duration of each frame was 10 ms, and a minimum of three states per segment had to be visited.

4.2.1.2 Resulting segmentation

The automatic phoneme segmentation was done for the 1244 sentences from the training set and the 1000 sentences from the test set. In the next step the syllable boundaries (derived via CELEX), the word boundaries, and the given prominence marks were added in the segmentation files. The segmentation of our example sentence *Hij heeft twee argumenten voor zijn stelling (He has two arguments for his proposition)* is shown in figure 4.1. For clarity, only the speech part of the whole

[3] The results in section 3.3.1 do suggest that, for instance, Negations require an independent class.

sound file is given above; some mouth noise in the beginning of the sentence as well as background noise at the end of the sentences has been cut out. The two debatable segmentation points that were mentioned above in section 4.2.1, (namely the possible junction of the two /t/'s between the words *heeft twee (has two)* and the /n/ deletion at the end of the word *argumenten (arguments)* can be checked in the speech signal. A /t/ de-gemination over word boundaries is not present in this specific sentence; the automatic segmentation even shows some silence between these two words. The /n/ deletion at the end of the word *argumenten* does not occur in this spoken sentence either, since it is clear that the marked schwa at the end of the word is exceptionally long, and the oscillogram indicates that there are two different speech sounds. Listening to the recording confirms that the /n/ was actually spoken here.

4.2.1.3 Accuracy of the automatic segmentation

The accuracy of the automatic segmentation is difficult to test; one option might be to compare it with a hand-segmentation of the same speech material. In automatic segmentation there are at least two types of error sources. Firstly, in our automatic segmentation we rely on the canonical pronunciation, which many speakers do not realize because of regional, gender type, and / or speaking style differences. Secondly, the segment boundaries set by the HMM recognizer can be inaccurate. The question now is, which deviations are acceptable, and how can we make an inventory of them. In order to bring more clarity the example sentence was segmented by hand. Figure 4.2 below shows a comparison between hand and automatic segmentation. A colleague independently performed the hand segmentation. The exact position of most boundaries differs somewhat and in one case an insertion has to be made. The /n/ at the end of the word *argumenten (arguments)* is definitely spoken and must be inserted for a correct segmentation. This error has consequences for the duration of the schwa and pollutes the measurements. There might be other structural errors in the automatic segmentation such as for the segmentation of vowel-like consonants (l, R, j). In automatic speech the alignment takes up a greater part of the adjacent vowel as compared to hand transcription. The /R/ in the word *argumenten* in figure 4.2 is an example of this.

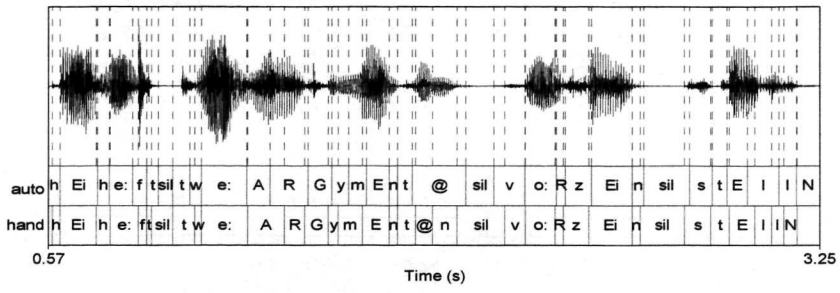


Figure 4.2: Hand-made segmentation (lower tier), in comparison with the automatic segmentation (upper tier) plus the oscillogram.

Although the automatic segmentation is far from perfect and structural errors cannot be excluded, we nevertheless decided to use the automatic segmentation for the following reasons: First, the main problem in this chapter is not only to find relevant acoustic features, but also to automatize the whole process of measuring acoustic features at the appropriate place in the speech signal, and to detect words with different degrees of prominence. Furthermore, hand segmentation is time consuming, which makes it difficult to segment large databases by hand. Therefore automatic segmentation is the only reasonable option. However, the more structural errors, such as the /n/ deletion, form a real problem, and it must be investigated to what degree these errors obscure reliable measurement of acoustic features. Lastly, the segmentation must also be done automatically in speech technology applications. To indicate the quality of automatic segmentation we can refer to Wang (1997), who compared the hand-labeled TIMIT database with the automatic segmentation (similar to the one used in the present study). He obtained a correct score of 86.9% within a 20 ms interval.

4.2.2 Acoustic correlates

In the next sections we will describe and discuss acoustic features of prominence. All the acoustical measurements and analyses were done with the PRAAT software package (Boersma & Weenink, 1996).

4.2.2.1 Unit selection

The sentences judged for prominence by ten listeners have been included in the acoustical measurements. Details of these 1244 sentences are described in section 2.2.4. This set is called the training set. The acquisition of the prominence marks is described and discussed in section 2.4. The clustering of the prominence marks as presented there, is used in the current chapter, so the prominence scale from 0 to 10 (originating from the 10 listeners) is reduced by hierarchical clustering techniques to a 4-point scale (see section 2.4.1.3), where III (score seven, eight, nine and ten) is the highest prominence class, followed by II (score three, four, five and six) and I (score one and two), with 0 (score zero) indicating no prominence at all. It is our goal, in this chapter, to find the acoustic features that are correlated with these prominence scales. These features allow us to distinguish between the four prominence categories (III, II, I, 0), or, to make the training classes more distinctive, between the two extremes of this scale (0 and III). For this purpose the 1244 sentences were used as the training set. The other set of 1000 sentences serves as a test set (see section 2.4.2). As testing the relevance of the acoustic correlates is basically the main topic of chapter 5, the data presented in the present chapter mainly concern the training set of 1244 sentences.
The listeners assigned prominence at the word level. Measurements at this word level might be advisable for F_0[4]. However, the word unit may be too large to

[4] In a pilot experiment some good results were achieved by taking the difference between maximum and minimum F_0 within one word. This measurement proved to be a useful feature for prominence; more details are described in Streefkerk et al. (1997).

measure duration and intensity. For this reason we decided to use the syllable as the basic unit to work with. If a word consists of several syllables it must be decided to which of the syllables the unique prominence label should automatically be allocated. We decided to use the criterion of lexical stress. Therefore, in a polysyllabic word the syllable that carries the lexical stress is the target one.
Figure 4.3 shows our example sentence, with the above-mentioned tiers of syllable boundaries and the original prominence marks of the listeners. The third tier shows the different prominence classes assigned to the syllables. As discussed above, for polysyllabic words the prominence class is assigned only to the lexically stressed syllable. The remaining syllables marked with '-' do not play a role whenever the so-called absolute features are applied, but they are actually used for normalizations.
In this approach it is assumed that features of prominence as defined by listeners are concentrated in the lexically stressed syllable of a word. The literature confirms that features of sentence accent are mainly concentrated in the lexically stressed syllable (Sluijter, 1995; 't Hart et al., 1990). Lexical stress as a property of the lexicon is at best an approximation of realized lexical stress. However, phenomena such as lexical stress shift, or unspoken lexical stress are not taken into account. If stress shift occurs then this indicates that we are measuring in the wrong syllable.
Linguists often define lexical stress via the content-word / function-word distinction. Content words do have lexical stress; function words do not. (More about this distinction is said in chapter 3.) Our choice means that all words must be included in the analyses; even Dutch articles such as *de (the, masculine)* and *het (the, neuter)*. In principle, such words can also be identified as prominent and since we are also interested in the degree of prominence, 'no prominence' is as interesting as 'prominence'. With this aim in mind, we had to be consistent and take all words into account, knowing that function words are generally not prominent (see chapter 3). This might even have consequences for certain measurements. For instance, the frequently occurring function word *de* contains a schwa. Features that apparently distinguish between prominent and non-prominent may actually distinguish between schwa and full vowels, for instance.
We have discussed several aspects of automatic segmentation, which is required to anchor the acoustical measurements, and to assign prominence to syllables and to select these syllables. The last point mainly dealt with the question as to which syllables play a role in further analyses.

4.2.2.2 F_0 Features

As already described in the introduction to this chapter, F_0 changes are the most important cue for prominence. But first we have to explain what we mean by F_0. Periodicity can directly be measured in the speech signal. What the periodicity (fundamental) of the signal is can be expressed in cycles per second (frequency). This is called the fundamental frequency, also named F_0. F_0 is measured in Hertz, which is the physical unit of cycles per second. As already mentioned, F_0 is closely related to the perception of pitch. The higher the F_0 values the higher the perception of pitch, but there is not a linear relationship between perceived pitch and F_0 measured in Hertz. To define the relation some people use a logarithmic, musical scale and measure F_0 in semitones. Other more psychophysical units are also

conceivable such as ERB (equivalent rectangular bandwidth) or bark (perceptual spectral frequency; approximation of the place on the basilar membrane). In the present study we chose semitones (st), because they come close to the perception scale of pitch and they are related to an easily interpretable musical interval scale. The relation of the acoustical frequency scale (Hz) and the logarithmic musical scale (st) relative to 100 Hz can be expressed as follows:

$$f = \frac{12\ln(F_0/100)}{\ln 2}$$

'f' is expressed in st, F_0 in Hz.
This means that a tone of 100 Hz is expressed as 0 st (this tone is 0 semitones above 100 Hz), and a tone of 200 Hz is expressed as 12 st. This tone is 12 semitones above 100 Hz, which is exactly 1 octave. A tone of 300 Hz is about 19 semitones above 100 Hz. An additional advantage of expressing F_0 in semitones is that the difference between two tones is independent of the reference value of 100 Hz. This independence allows F_0 ranges (difference between two F_0 values) to be directly comparable.
All these explanations concern a steady F_0, but in speech we have to deal with F_0 changes, which are perceived as a melody. If voiceless parts appear, for instance, in an /s/, no pitch can be measured. An F_0-curve thus contains gaps and discontinuities, but these gaps are generally not perceived as such. Perception closely resembles a continuous contour i.e. the speech melody. We have to approximate the perceived melody from the measurable parts of the F_0-curve. Striking changes in the F_0-curve may be used to mark prominence. These changes have to catch the ear, but other factors influence the F_0-curve as well (such as male / female differences, declination, and boundary tones) and may disturb a direct feature extraction. (An evaluation of a

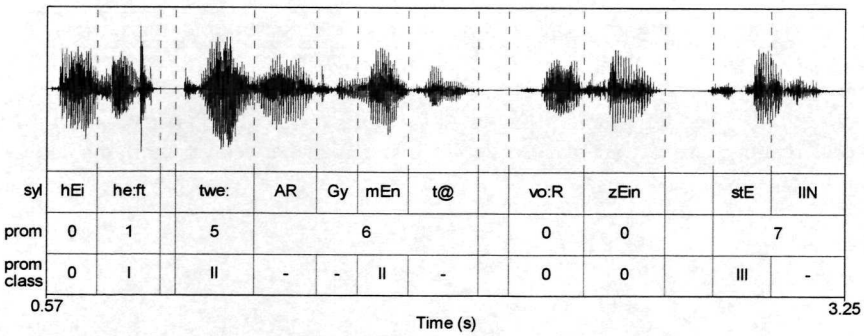


Figure 4.3: Oscillogram of the example sentence *Hij heeft twee argumenten voor zijn stelling*, with from top to bottom the segmentation at the syllable level ('syl'), then the prominence marks of the listeners at the word level ('prom'), and the selection of the syllables used with the prominence degree levels (0, I, II, III) that go with these syllables ('prom class').

parameterized F_0-curve is described in Heuft et al. 1996.) Simple interpolation and smoothing techniques approximate the perception. Methods must be found to extract features that catch the striking changes in the F_0-curve. These methods must also be corrected for the disturbing factors, such as declination.
Having explained some terms concerning the acoustical phenomena of F_0 and its perceptual correlates pitch, we must now turn to the analyses of the actual F_0 measurements in our sentences. The main topic of the following sections is the automatic extraction of the relevant features from the whimsical F_0-curve.

4.2.2.2.1 Measuring F_0

For the F_0 measurements a periodicity detection based on an accurate autocorrelation method is used (Boersma & Weenink, 1996), with time steps of 0.01 seconds. The maximum frequency found is set at 600 Hz (it was unusual for a woman's voice to go beyond this value). The minimum is set at 50 Hz (it was also unusual for a man's voice to go below this level). The unprocessed F_0-curve (see figure 4.4, graph 1) is first automatically corrected for octave jumps. This correction step is not without problems. In Dutch, the range of the whole curve per utterance per speaker is normally within one octave, so measurements outside this range are usually not expected. Therefore, we implemented a correction step (see figure 4.4, graph 2), but sometimes this correction step introduced new errors. This actually happened in our example sentence. Through repeated listening it was confirmed that the sentence onset *hij heeft... (he has)* is actually as high as presented in the original upper graph of figure 4.4. The resulting F_0-curve, of course, shows gaps for the voiceless parts in the speech signal. The curve per sentence is interpolated by a simple linear fit in order to get a continuous curve (see figure 4.4, graph 3). As already mentioned, we believe that human perception is closer to this continuous contour. This continuous contour has the advantage that measurements at the syllable level are now possible, and that a value can be extracted for each syllable.

This interpolated contour is again smoothed to eliminate small variations in the curve (see figure 4.4, graph 4). Such small variations may influence the feature extraction negatively as local fluctuations may not express the more global changes in the curve. The last graph of figure 4.4 shows the actual F_0-curve from which several features are extracted in order to distinguish between prominent and non-prominent words.

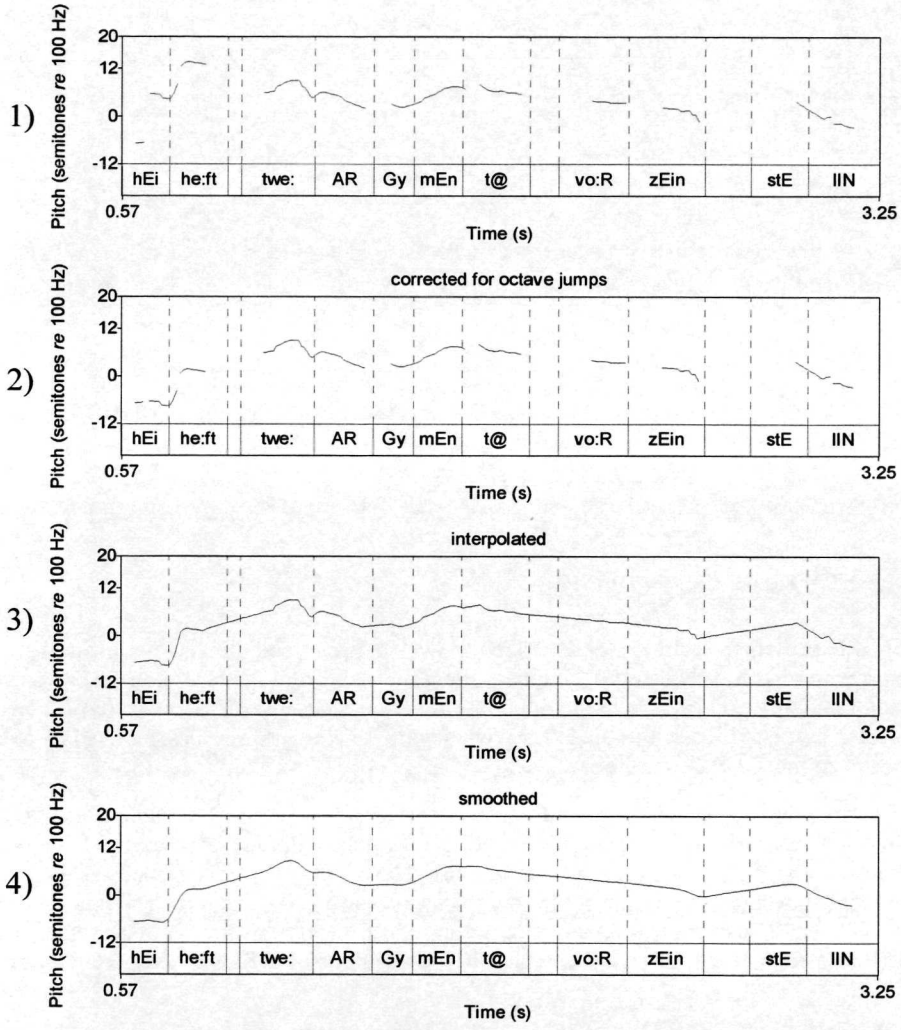


Figure 4.4: F_0-curves of our example sentence *Hij heeft twee argumenten voor zijn stelling*. The first graph displays the unprocessed F_0-curve; in the second graph the octave jumps are corrected; in the third graph the F_0-curve is linearly interpolated in order to make the curve continuous; and in the fourth graph the curve is smoothed to get rid of small local changes which could disturb the automatic extraction of F_0 features.

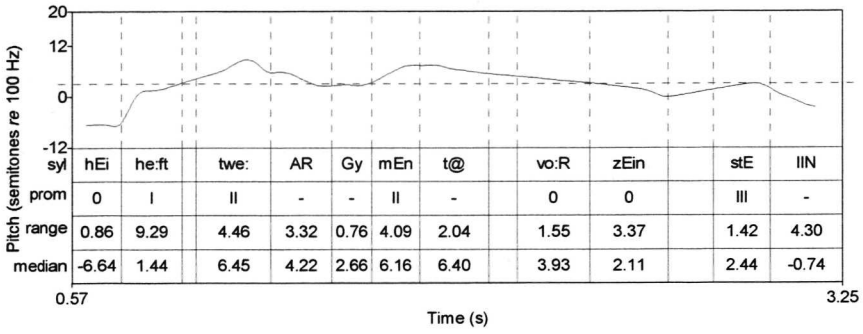


Figure 4.5: The F_0-curve with the syllable segmentation and the prominent classes. The two lowest tiers show the range and the median F_0 per syllable. A dotted horizontal line through the curve indicates the overall median.

4.2.2.2.2 Extracting F_0 features

In the previous section we described how we processed the F_0-curve so that it corresponded more closely to human perception, and how features can be extracted automatically for each syllable. A continuous curve is useful because only then a vector can be filled with values for each segment. These values can be used as cues for prominence classification.

The median F_0-value per syllable as well as the F_0 range per syllable (difference between maximum and minimum of the target syllable) is measured on the continuous pitch curve (see section above). The values of F_0 range per syllable are directly interpretable. Because these values express a difference, the reference value of 100 Hz is no longer relevant. E.g. a range of 3 semitones means that the interval of these tones is a minor third, which is independent of different pitch heights of a voice.

However, the syllable-based median values cannot directly be compared to each other. These values can only be compared within one sentence. So, in order to compare a sentence spoken by a male with a sentence spoken by a female, gender normalizations must be made. This was implemented by applying an utterance-based normalization by subtracting the overall median F_0-value of each sentence. These resulting corrected median values thus express the median deviation per syllable from the median value of the whole sentence.

The range and the median F_0 values are shown per syllable for our example sentence in figure 4.5. For instance, on the lexically stressed syllable of the word *argumenten* /mEn/, a movement is realized with a range of 4.09 st, whereas on the word *voor* /vo:R/ the curve has a range of 1.55 st. The median per syllable is measured to give an indication of the overall height of a syllable. Since for this example we only

compare within one sentence, we show here the uncorrected values. The syllable *twee* has the highest median F_0 value of all the syllables, followed by the last two syllables of the word *argumenten*, which have median values of 6.16 and 6.40 st.
Is it possible to distinguish which words of our example sentence are prominent by only using these two features (F_0 range, F_0 median per syllable)? The extracted features then have to distinguish between the different prominence classes. As decided in a previous section, for these direct analyses of polysyllabic words, only the lexically stressed syllables are taken into account as measurement domain. The discarded syllables are marked with a dash (-) in the second tier of figure 4.5. A high F_0 range value is an indication of prominence. Looking at this sentence, the candidates are *heeft* 9.29 st, *twee* 4.46 st, the syllable /mEn/ 4.09 st and perhaps *zijn* with 3.37 st. A syllable with a high F_0 median can also be an indication of prominence. Taking the median per syllable we have the following candidates: *twee* 6.45 st and the syllable /mEn/ 6.16 st. On the condition that both features must be high, we may conclude that the word *twee* and the word *argumenten* are the most prominent ones in this sentence. But we have a problem: the word *stelling* is completely discarded by these features, whereas the word *heeft* could wrongly be classified as being prominent, as it has a high F_0 range value. The wrong classification of the word *heeft* may be caused by the F_0 measurement; such mistakes are plausible. But the fact that the present F_0 features do not support the prominent classification of the final word *stelling*, could constitute a structural problem. Bulyko & Ostendorf (1999) also observe that prominence occurring late in the phrase is harder to predict. However, there is one striking pitch movement relative to the environment of this word at the end of the sentence. The median values per syllable do not help either to mark this syllable as prominent, which means that we have a similar problem with this feature. At the end of the sentence the overall height of the F_0-curve goes down. Therefore median values per syllable are generally lower at the end of the sentence than in the beginning. So the next step could be to correct for these errors by taking into account the declination line of a sentence and the direct environment. Attempts to correct for the declination line failed however. In a pilot study we subtracted the error-prone linear fit (an approximation of the declination line) per sentence for a subset of 500 sentences. With this corrected F_0-curve the same features, namely the F_0 range per syllable, and the corrected median per syllable were calculated. Unfortunately, the features corrected for the declination line did not increase the ability to discriminate (Streefkerk et al., 1999 a).
In order to get an overall impression of the pitch movement in a word we also decided to measure the F_0 range per word. For our example sentence this results in the following F_0 range per word:/hEi/ 0.86, /he:ft/ 9.29, /twe:/ 4.46, /ARGymEnt@/ 4.41, /vo:R/ 1.55, /zEin/ 3.37 and /stELIN/ 4.85. Of course only the values for the polysyllabic word *argumenten* and *stelling* differ. Looking at the last word of the example sentence it is clear that taking the F_0 range per word more closely expresses the spoken pitch movement than the same measurement per syllable.

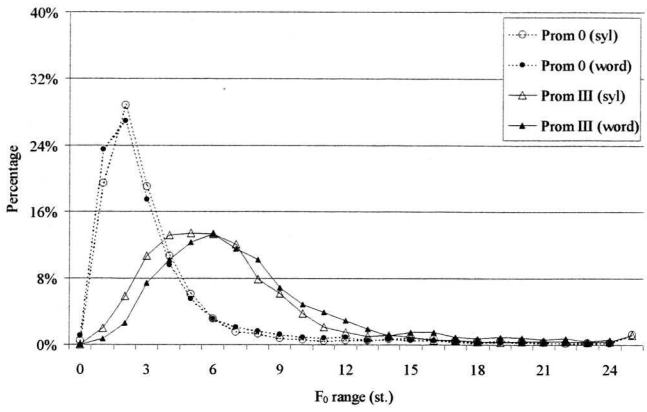


Figure 4.6: Distribution of the prominence class III (total number of 2872 words) and prominence class 0 (total number 5977) for the feature F_0 range per syllable and F_0 range per word, both in semitones.

Next, let us look at some overall distributions for these three features (F_0 range and median F_0 per syllable, and F_0 range per word). Figures 4.6 and 4.7 show histograms with the distributions for the extremes of the prominence classes: the non-prominent class (0) and the most prominent class (III). For words of more than one syllable the value of the lexically stressed syllable is shown, as explained in a previous section 4.3. We chose to show only the extremes 0 and III of the prominence scale for simplicity's sake. The total numbers for these two distributions differ (prominence class III, 2872, and prominence class 0, 5977[5]) so, in order to make these distributions comparable, we express them as percentages so that the surface under the curves is the same, see figure 4.6. For the F_0 range per word we can see that the distribution for prominence class III is more shifted to the right than for the distribution of the F_0 range per syllable. Whereas the F_0 range per word distribution and the F_0 range per syllable more or less overlap. The further apart the 0 and III distributions, the more the given feature can distinguish between these two prominence classes. Looking at these two distributions a distinction can definitely be made, but how accurate this distinction actually is, will be discussed in chapter 5. To give an overall impression of the ability to discriminate we give the amount of overlap. The overlap (the area these two distributions have in common) is 49% for the F_0 range per syllable. The amount of non-overlap between these features indicates their ability to discriminate between prominence (III) and non-prominence (0). For the F_0 range per word this area decreases and the amount of overlap is reduced to 42%, see figure 4.6.

[5] In chapter 2 the total number of 0 marks is 5950 (see table 2.6). Here we have a total number of 5977, because some stutters or repetitions are segmented as separate words. These 27 stutters and repetitions are marked with a 0, so that they can join the further analyses.

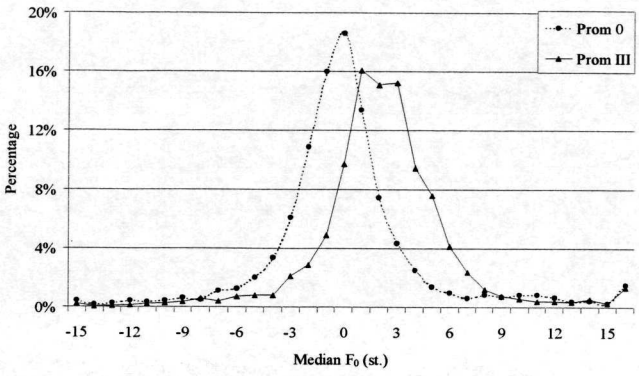


Figure 4.7: Distribution of the prominence class III (total number 2872) and prominence class 0 (total number 5977) for the feature median F_0 per syllable, corrected for the overall F_0 median per sentence in semitones.

In figure 4.7 the median F_0 per syllable (of which the overall median per sentence is subtracted) is displayed. The two distributions for prominence class 0 and prominence class III are not as clearly separated as in figure 4.6, the amount of overlap being 59%. But in combination with other features this feature may also contribute to differentiate between prominent and non-prominent words.

4.2.2.3 Duration features

Acoustic features concerning duration are directly available from the segment markers; so no further acoustical analysis of the speech signal is necessary. The duration of speech units e.g. vowels, consonants, syllables and words can be extracted directly from the segmentation file and can be used for analysing the effect of prominence on the duration. However, such durations must be expressed relative to their environment. Generally, the duration of speech units is influenced by other factors such as style, the speaker, the speaking rate, and the intrinsic duration of speech segments. The position of a segment in a word or in a sentence (final lengthening) also influences the duration, but in this research no further attention is given to the co-influence of these effects and prominence.
However, normalization might be necessary for the previously mentioned factors. In the following sections this normalization is investigated in detail. It is not sure, whether or not the effect dedicated to prominence is clearer after normalization. The vowel and syllable durations of the example sentence are shown in figure 4.8.

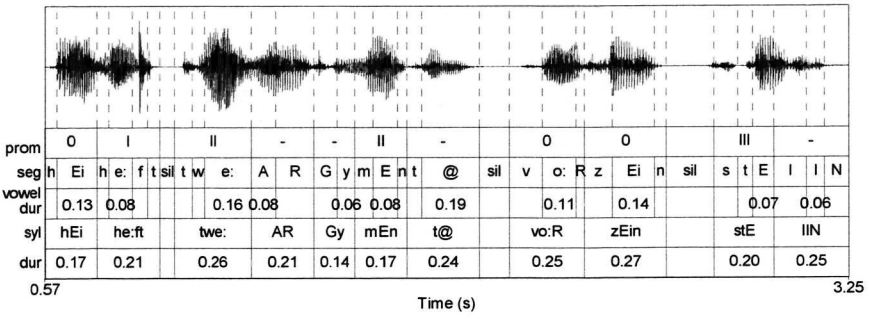


Figure 4.8: The oscillogram of the example sentence *Hij heeft twee argumenten voor zijn stelling* is given, followed by the prominence classes, the phoneme and syllables segmentation plus their durations.

One might argue that the longer a speech unit the more prominently it is perceived. However, in this example sentence several counter examples can be found. First of all the longest vowels and syllables are not those occurring in prominent words. The vowels from the lexically stressed syllables in prominent words belong to the short vowels category. The duration of the vowel /E/ in the word *argumenten* is 0.08 s and of /E/ in the word *stelling* is 0.07 s. However, the monosyllabic word *twee* also has a long duration of 0.26 s, of which 0.16 s belongs to the vowel (see figure 4.8). The example sentence indicates that normalization for intrinsic vowel duration might be useful, but let us first see how far we get with raw data such as vowel and syllable duration.

Since the phonemes of each sentence were automatically segmented, the duration of each segment is determined by the HMM-recognizer. The minimum duration of each vowel is thus 0.03 s, as the duration of each frame was fixed to 0.01 s and a

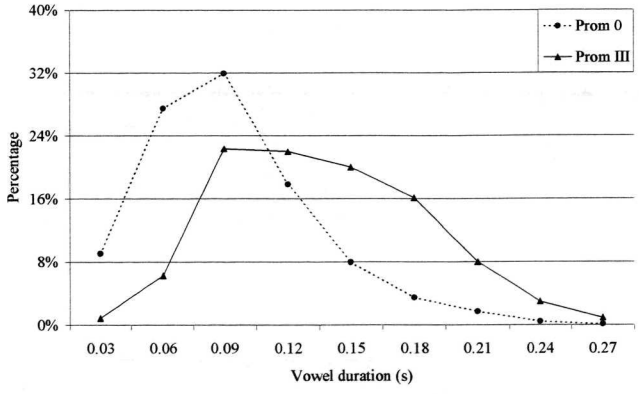


Figure 4.9: Distribution of the prominence class III (total number 2872) and prominence class 0 (total number 5977) for the duration per vowel concerned.

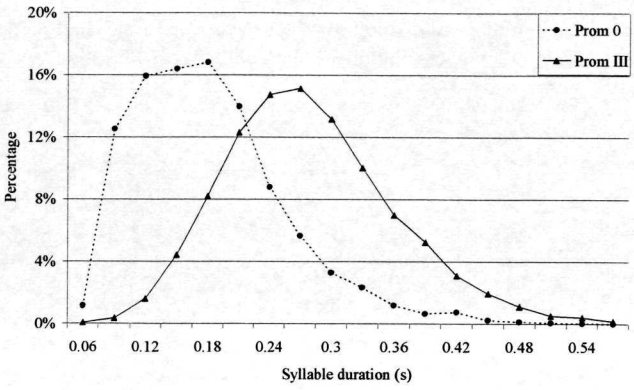


Figure 4.10: Distribution of prominence class III (total number 2872) and class 0 (total number 5977) for the duration per syllable.

minimum of three states per vowel must be visited (see section 4.2.1.1). This explains why in figure 4.9 the duration distributions start at 0.03 s. Figure 4.9 shows the distribution of vowel duration both for prominence classes 0 and III. The amount of overlap is 61%, which is 20% higher than for the F_0 range per word, but a certain distinction based on the feature of vowel duration can still be made.

Portele & Heuft (1997) found that syllable duration is a useful feature of prominence. For our syllable duration data (see section 4.3) the duration for the two extremes in the prominence classes are shown in figure 4.10. These two distributions are more separate from each other than the distributions for the vowel durations. The overlap of these two distributions is 50%, thus 11% less than for the distributions of vowel duration. To distinguish between non-prominence (0) and prominence (III) it seems therefore that the syllable duration is a more useful feature than vowel duration. Possible segmentation errors lend more weight to the small vowel unit than to the larger syllable unit. Often function words contain fewer clustered consonants, which decreases their duration.

In distributions concerning the vowel durations, the short prominent vowels possibly overlap with the long non-prominent vowels. It is thus possible that corrections for intrinsic vowel duration will influence the discriminative power. Table 4.1 gives the mean vowel duration and its standard deviations, and the frequency of occurrence of each vowel. It is generally known that diphthongs are the longest (Au = 0.17 s, Ei = 0.134 s, and 9y = 0.157 s expressed in mean values) (e.g. Koopmans- van Beinum, 1980). Our long vowels (indicated with an ':') have indeed a rather long duration too, whereas the so-called mid-long vowels are shorter (see table 4.1) and seem to fit into the short vowels category. The different frequency of occurrence of the various vowel classes in this Polyphone corpus is an interesting phenomenon. The schwa occurs most frequently whereas most diphthongs and some long vowels are rather rare. The accompanying distribution is shown in figure 4.11. This histogram shows the vowel duration distribution for the schwa as well as for all other vowels

Table 4.1: The number of different vowels occurring in the training data (1244 sentences with all vowels included) with their mean duration and standard deviation in seconds.

	Vowel	Num	Mean	Std. Dev.
Schwa	@	7384	0.060	0.041
Diphthongs	Au	398	0.170	0.044
	Ei	1012	0.134	0.051
	9y	287	0.157	0.040
Short	A	1998	0.082	0.033
	E	1574	0.082	0.036
	I	1605	0.082	0.037
	O	1360	0.092	0.038
	Y	393	0.068	0.031
Long	Q:	267	0.140	0.042
	a:	1588	0.134	0.052
	e:	1446	0.129	0.045
	o:	1165	0.117	0.051
Mid-long	u	460	0.090	0.039
	i	1238	0.086	0.038
	y	321	0.102	0.052
Total	-	22496	-	-

combined. It is striking that the schwas are short, but the tail of the distribution is very long and the other vowels are generally longer. This might partly be caused by lexical stress, as schwas do not occur in a lexically stressed position. For our classification only the lexically stressed syllables of polysyllabic words are taken into account. This reduces the number of schwas, but they are still present in the monosyllabic function words.

Corrections for 'intrinsic' vowel duration were carried out in the following way:

$$\tau = \frac{d - \mu}{\sigma}$$

where d is the duration of the given vowel, and μ and σ are the mean and the standard deviation of the corresponding class, respectively. The result is that the vowel durations are expressed in z-scores, which have the property that the mean is 0 and the standard deviation is 1.

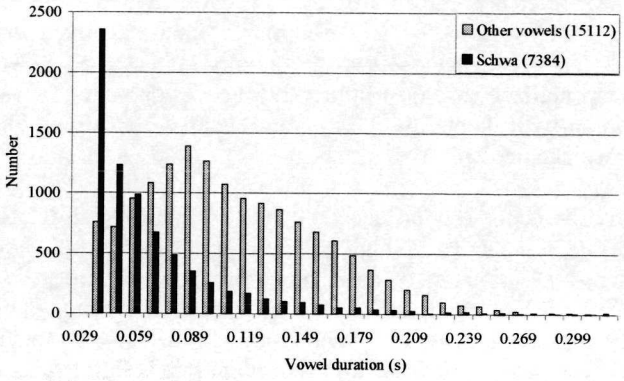


Figure 4.11: Two separate distributions of the vowel duration for schwa and other vowels.

Because the duration measurements are discrete (the minimum duration of a vowel is 0.03 s and the increase is in 0.01 s steps), display artefacts arise. This is visible in figure 4.12. Because of the discrete durations the normalized data are also discrete. This means that for instance a schwa of 0.03 s always corresponds to a normalized value of -0.73 ((0.03 - 0.06) / 0.041), whereas a schwa duration of 0.04 s always gives the normalized value of -0.49, and so on. But these steps are different for each vowel class. This effect makes it difficult to put the data into a smoothed histogram. In figure 4.12 we have chosen the bin steps in such a way that they correspond to the discrete steps of the schwa. This implies that there are no values of that given vowel between the two steps. Despite the high frequency of occurrence of the schwa the irregularity effect is still present. The irregularities at one quarter and three quarters of the histograms are due to the fact that the normalized durations of the short vowels do not occur between +0.98 and +1.22 and between -0.98 and -1.22.

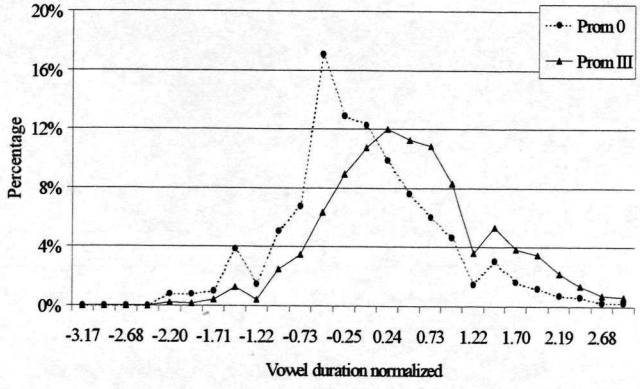


Figure 4.12: The distribution of vowel duration normalized for the intrinsic vowel duration separate for the two extremes of the prominence classes (0 and III).

Because of this effect it is rather difficult to judge whether or not the normalized duration is a useful feature for the prominent / non-prominent distinction. The amount of overlap between these two histograms increases to 72%, compared to 61% in the unnormalized case (see figure 4.9). Feeding a neural network with these features as input will hopefully give more details concerning their ability to discriminate (see chapter 5).

The vowel duration is not only influenced by the intrinsic vowel duration but also by the speaking rate. Generally, a high speaking rate shortens the duration of all segments. There are various methods of measuring speaking rate. A simple method is to count the number of segments in 1 second of speech, leaving pauses out. Such a method highly depends on which segments occur in the given stretch of speech. If only intrinsically long segments occur, the speaking rate is falsely classified as slow.

The method we used (Wang, 1997) corrects for these effects, and leaves out pauses. The formula given below defines the sentence speaking rate (r) as the average normalized segment duration per sentence whereas τ denotes the normalized segment duration.

$$r = \frac{1}{N}\sum_{i=1}^{N}\tau_i \qquad \tau = \frac{d-\mu}{\sigma}$$

If the sentence speaking rate (r) is 0, the speaking rate is average; negative values of r indicate that the speaking rate is high (because the segment durations are on average shorter than the mean durations of each segment). Positive values indicate a slow speaking rate. Figure 4.13 shows the distribution of the sentence speaking rate of the 1,244 sentences. The rates are equally distributed around 0.

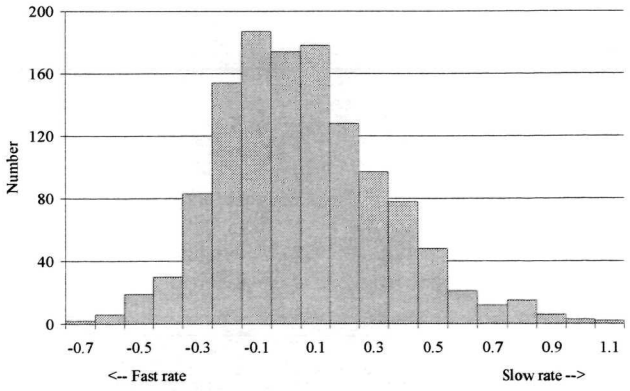


Figure 4.13: The distribution of the sentence speaking rate of the 1244 sentences.

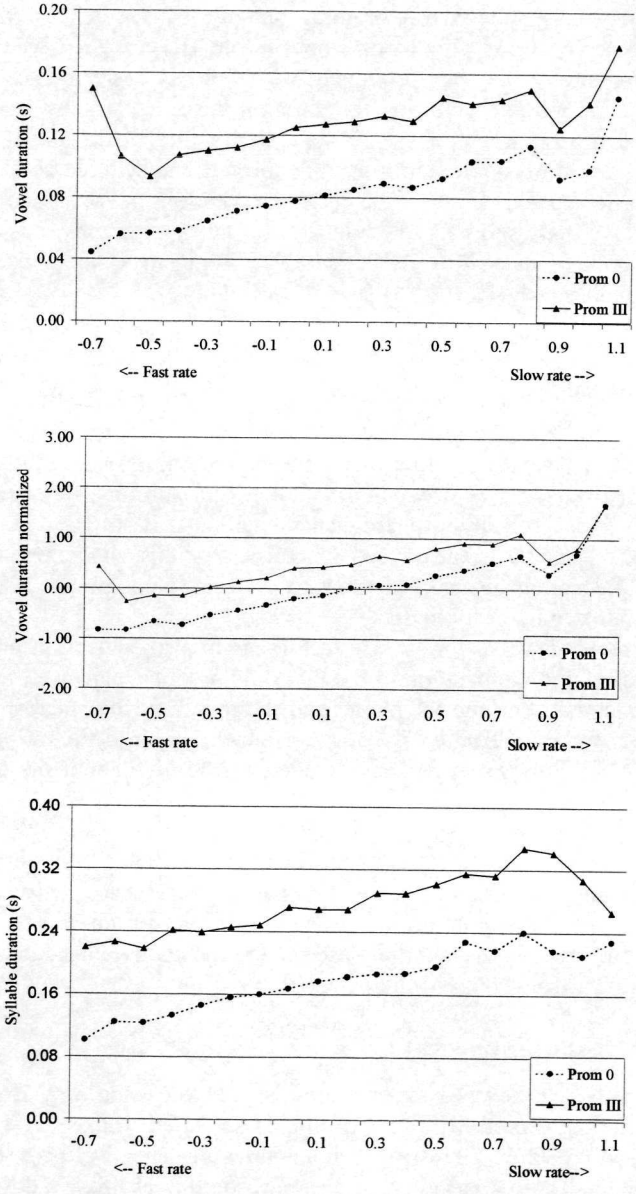


Figure 4.14: Average vowel duration, average normalized vowel duration, and average syllable duration as a function of the sentence speaking rate. In each graph the two curves represent the most (scale III) and the least (scale 0) prominent vowels or syllables, respectively.

Figure 4.14 shows a distribution of the mean vowel duration as a function of the sentence speaking rate in the training set of 1244 sentences. Since the vowels belonging to the most prominent words (scale III) are displayed separately from those belonging to the least prominent words, it can easily be seen that the former are substantially longer and that the duration increases at slower sentence speaking rates. The lowest graph of figure 4.14 shows the relationship between sentence speaking rate and syllable duration. As expected, the syllable duration also shortens with a faster speaking rate. The average durations at the edge of the sentence speaking rate scale show large deviations for all three graphs; this is because of the small amount of data. Vowels or syllables from only two or three sentences are involved at these edges.

4.2.2.4 Intensity features

For the features concerning intensity we have to consider also the relationship between the physical unit 'intensity' and its psychophysical counterpart 'loudness'. Intensity (dB) is a physical unit relative to the auditory threshold. The auditory system is most sensitive to frequencies around 1000 Hz. The loudness level (measured in sones) is introduced to correct for this characteristic. However, the loudness perception of complex tones is even more complicated. So we decided to use as a first step intensity in dB.
Intensity normalizations for several factors might also be useful. Such factors are the type of vowel and the position of the word in the sentence as well as lexical features. The distance between the telephone and the mouth of the speaker is another factor influencing the overall intensity of a sentence. But since no calibration signal was applied in the Polyphone database we have little idea about the absolute intensity level.

Generally, most of the differences in intensity for being prominent or not can be found in the vowels. For the consonants, differences between the different consonants are too large to allow for an overall comparison. For the vowels there are intrinsic differences; this is mainly ascribed to the open / closed distinction of vowel, and secondly to the front / back distinction.

4.2.2.4.1 Measuring intensity

The intensity per vowel was measured in the following way. In the signal each vowel was segmented with a two-timed, so-called Kaiser 2 window (PRAAT, Boersma & Weenink, 1996) at the boundaries of the vowel given by the automatic segmentation. The overall intensity was measured over this windowed vowel signal. The individual results for the vowels in the example sentence are given on the bottom tier of figure 4.15. For the whole sentence the intensity curve is displayed in this figure. The individual values do not correspond in an absolute sense with the intensity scale of the intensity curve in figure 4.15, because the loudness per segment is computed in a different way, mainly due to the cutting of the Kaiser 2 window. The syllables bearing no lexical stress are actually spoken with less

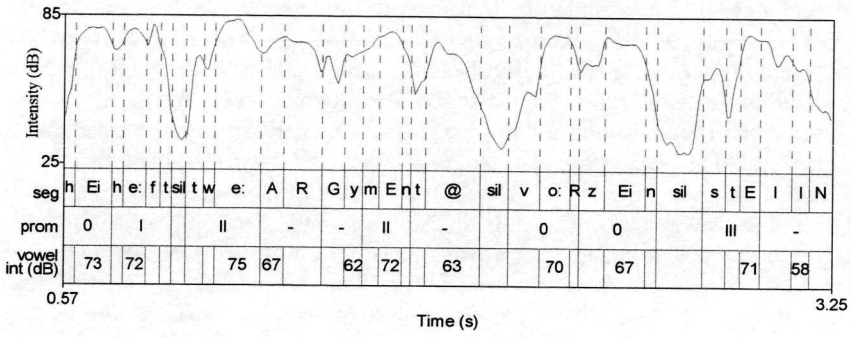


Figure 4.15: The intensity curve (dB) of the example sentence *Hij heeft twee argumenten voor zijn stelling*. In the upper tier the phonemes are presented, in the middle tier the prominence classes and in the last tier the intensities (dB) of the vowels.

intensity than their lexically stressed counterparts. As far as prominence is concerned, we expect that if a word consists of more than one syllable, the lexically stressed syllable will have the highest intensity. The vowels in the syllables of the most prominent words (/twe:/, /mEnt/ and /stEl/) do indeed show high intensities (75 dB, 72 dB, and 71 dB, respectively). But other vowels do so as well, such as the first two vowels of the sentence (/Ei/ and /e:/), with an intensity of 73 dB and 72 dB, respectively. Moreover, the intensity of the vowel /o:/ in the non-prominent word *voor* is still rather high. But the vowels in the less prominent words may have a higher intrinsic intensity or contain a diphthong, which may cause other complications.

To obtain an overall impression of the discriminative power of the loudness features of the vowels, two distributions are plotted in figure 4.16. As an overall normalization the overall intensity of the sentence is subtracted (to each value 80 dB

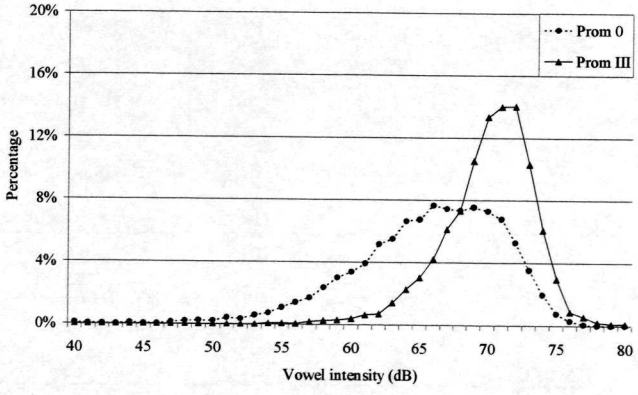


Figure 4.16: The distributions of vowel intensity (dB) spliced for the most (III) and the least (0) prominent vowels.

is added in order to get realistic values). We observe that the vowels in words in prominence class III are generally spoken with higher intensity than the vowels in prominence class 0. This figure only shows that the effect of prominence is noticeable and that the intensity of vowels is a useful cue in itself, even without any normalization for vowel type. The overlap of these two curves is 57%.
We discuss the normalization for intrinsic vowel intensity in this section. Generally it is suggested that open vowels are louder than closed vowels, and that front vowels are louder than back vowels. Table 4.2 gives the average intensity and the standard deviation per vowel class. Looking at the distinction between, open and closed vowels and at front, mid and back vowels, it is not confirmed that the open vowels or the front vowels are generally louder than their counter parts. The differences between the vowel classes itself are also not so striking, however, a normalization per vowel class, according to the formula given below, might help to increase the power to discriminate.

Table 4.2: The number of different vowels occurring in the training data (1244 sentences with all vowels included) and their mean intensity and the standard deviation.

		Vowel	N	Mean (dB)	Std. Dev. (dB)
		@	7384	73.45	5.89
Front	closed	i:	1238	74.28	4.71
	closed	I	1605	75.21	5.24
	closed	e:	1446	78.75	3.45
	closed	Ei	1012	78.45	3.53
	open	E	1574	78.08	4.43
Mid	closed	y:	321	75.09	4.84
	closed	Y	393	77.84	4.14
	closed	Q:	269	79.46	3.55
	closed	9y	284	79.02	3.24
	open	A	1998	78.09	4.02
	open	a:	1588	78.20	3.45
Back	closed	u	460	76.41	4.07
	open	Au	399	79.57	3.01
	open	O	1360	76.74	4.68
	closed	o:	1165	78.16	4.03
Total			22496	-	-

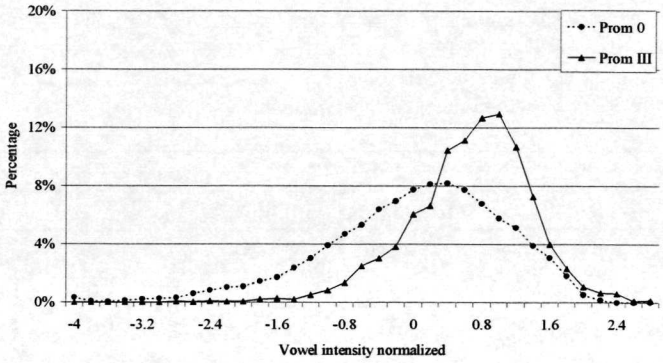


Figure 4.17: The distribution of vowel intensity normalized for the intrinsic vowel intensity, displayed separately for the two extremes of the prominence classes (0 and III).

$$\iota = \frac{I - \mu}{\sigma}$$

where I is the intensity of the given vowel, with the average μ and the standard deviation σ. As described before, the vowel intensities are corrected for sentence intensity, and 80 dB is added to these values in order to make them more realistic. The results of this intensity normalization are shown in figure 4.17. The two distributions of the two prominence class extremes show even greater overlap (68%) than in the unnormalized version (57%). Based on these data it can be concluded that this normalization harms the power to discriminate.

Another possibility is to use spectral slope, which however, is generally considered to be a rather difficult feature to use. A study based on a large variety of speech material shows little success with the spectral slope feature (van Kuijk & Boves, 1999). Contrary, the study of Fant & Kruckenberg (1999) concerned with well-recorded prose reading in Swedish, shows a very high correlation r = 0.93 between spectral tilt (SPLH-SPL) and prominence (marked for every syllable on a semi-continuous 36-point scale).

Table 4.3: Summary of the various acoustic features that have been discussed in this chapter. The amount of overlap between prominence classes 0 and III is indicated, as well as the ranking value based on that overlap. For both extreme prominence classes (0 and III) the mean value and its standard deviation per feature are presented in the rightmost columns.

			Prom class 0		Prom class III	
Feature	Overlap(%)	Ranking	Mean	Std. Dev	Mean	Std. Dev
F_0 range per syllable	49	2	3.59	4.79	6.26	4.46
Median F_0 per syllable	59	5	-0.30	4.69	1.82	4.09
F_0 range per word	42	1	3.25	4.07	7.48	4.84
Vowel duration	61	6	0.08	0.04	0.13	0.05
Vowel duration normalized	72	8	-0.16	0.88	0.34	0.95
Syllable duration	50	3	0.17	0.08	0.27	0.09
Vowel intensity	57	4	75.87	4.96	80.07	3.02
Vowel intensity normalized	68	7	-0.09	1.06	0.60	0.71

4.3 Summary and conclusion

In the current chapter several acoustic features have been discussed and their potential ability to discriminate between prominent and non-prominent words is shown. In terms of overlap table 4.3 summarizes these features according to their amount of overlap, as determined by the distributions for the two extremes of the prominence scale (0, III). The mean values per class of the given feature and their standard deviation are also presented. A ranking of the individual features is given, based on the amount of overlap. A small overlap indicates a higher ability to discriminate. The table shows that F_0 range per syllable and per word are the most promising features, followed by syllable duration and vowel intensity. Normalizations at utterance level do not show the expected decrease in overlap and the standard deviation is very high.

A more detailed analysis of the individual and discriminatory power of the features will be carried out in chapter 5. Combining acoustic features will hopefully increase the classification results. The correlation between features may on the other hand not increase the classification capability.

5

NEURAL NET CLASSIFICATION OF PROMINENCE WITH ACOUSTIC INPUT FEATURES[1]

Abstract

The main topic of this chapter is the classification of word prominence, exclusively based on acoustic input. Prominence is classified by means of feed-forward networks. This chapter includes a brief description of such neural networks. The input is chosen from the acoustic correlates as described in chapter 4. In that chapter two applications were suggested, namely a sentence disambiguator and a word prominence indicator. The consequences of such applications in terms of training factors are examined in this chapter. We obtained a prominent / non-prominent classification accuracy of 82% for the development test set and 79% for the independent test set.

[1] Parts of this chapter were published in Streefkerk et al. (1998), Streefkerk et al. (1999 a), Streefkerk et al. (1999 b), and Streefkerk et al. (2001).

5.1 Introduction

In this chapter we deal with the classification of prominence based exclusively on acoustic input features. The subsections of this introduction give a general description of a neural network and two examples of possible applications for prominence classification. These applications bring us closer to the question: what do we want the network to do. Simple feed-forward networks are used to recognize prominence with selective data. These exemplary data are presented to the neural net during the training phase and indicate which features are especially important for achieving correct classification. The input pattern consists of characteristic acoustic features, as described in chapter 4. Based on such characteristic input patterns, a trained neural net predicts whether or not words are prominent. The advantage of neural networks is that no specific knowledge has to be expressed in rules; instead, the knowledge is based on the training data sample. Other classification techniques are possible, for instance, a linear discriminant analysis (LDA). In such an analysis, however, only linear relationships can be found. Representing complex relationships may be needed for our classification problem. Examples for complex data classification with neural networks are described in Weenink (1991) and in Lippmann (1987). For our analyses we try to keep the topology of the net as simple as possible.

5.1.1 How feed-forward networks work

Feed-forward networks consist of units (nodes) and activation functions. The basic units of neural networks are the nodes; several nodes are grouped into layers (see figure 5.1). A learning algorithm allows the neural network to learn a certain task by adjusting the weights. The layer(s) between the input and the output layer are called hidden layer(s), (see figure 5.1). The formula below expresses the relationship of the output with the output of nodes in a previous layer (x_i), weights associated with the connections (w_i) and a threshold (θ) of the node. With these variables the output (y) is calculated, via an activation function (f), for instance a sigmoid function. Each node converts the pattern of incoming activities into one single activity. This single activity ('output') is passed on to the other connected nodes in the next higher layer. The activation function typically falls into one of three categories:

- a linear function (the output is proportional to the total weighed input),
- a threshold function (the output is set at one of two levels, depending on whether the total input is greater than or less than some threshold value), or
- sigmoid functions (the outgoing activity varies non-linearly with the input (weighted)).

Generally the activation function can be expressed as:

$$y = f\left(\sum_{i=1}^{N} w_i x_i - \theta\right)$$

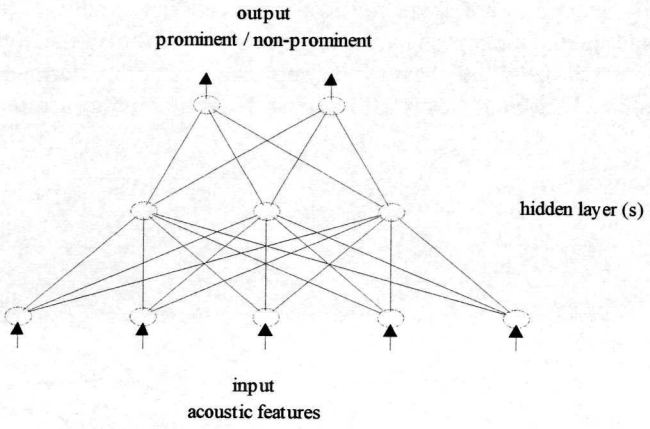


Figure 5.1: A possible topology of a neural network.

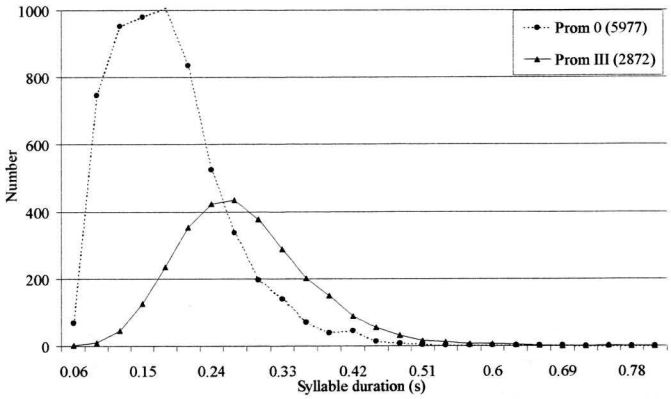
The behavior of a neural net depends on the weights (w_i) of the connections between the nodes, on the threshold (θ) of the node, and on the activation function (f). The weights and the threshold values can be adjusted during training; the activation function is usually fixed during training.

5.1.1.1 General training procedure

The main steps of the supervised learning process of the net can be described as follows: The acoustic features are presented at the input layer. The values of the features are sent via the weighed connections of this input layer and via weighed connections of one or more hidden layers to the output layer. The computed output is compared with the desired output values. Based on this comparison, the difference is calculated and the weights and the threshold values are subsequently adjusted (e.g. by the back-propagation algorithm, Rumelhart et al., 1986).
In the back-propagation algorithm the weights are adjusted per training pair (input features plus desired output) by using a feedback step. This makes the adjustment rather sensitive to the sequence in which the training pairs are presented to the net. A more sophisticated algorithm is the conjugative gradient method (Press et al., 1992). This training algorithm calculates the difference between the desired output and the calculated output for the whole training set, and then starts to adjust the weights and the threshold of each node. This whole procedure is called an iteration step. Such a training procedure makes training less sensitive to local minima. In this study we use the more sophisticated conjugative gradient method.

If enough examples are presented to the net, the net is able to generalize over the various characteristic input features. For new (unseen) data this trained net is able to predict which label belongs to the presented input pattern. The net bases its knowledge on the examples it has seen before in the training session.

a)



b)

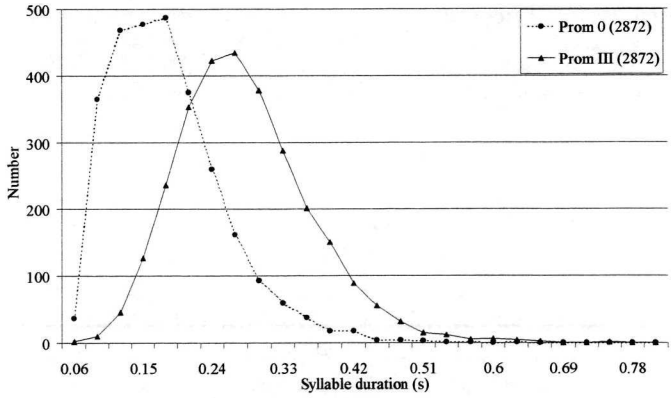


Figure 5.2: Two different distributions of the syllable duration for prominence class 0 and III. Graph a) displays the distribution based on the actual number of occurrences as found in the set of 1244 sentences. Graph b) displays two distributions of which the sizes have been made equal by random selection.

5.1.2 Distributions of prominence classes and the relationship to applications

Two applications of prominence classification have already been mentioned in chapter 4. One application is a prominence-indicator that measures the amount of prominence that each word carries. This classification is generally based on a biased distribution of prominent and non-prominent words (unequal numbers of prominent and non-prominent words). See figure 5.2a for an example of such a biased distribution concerning syllable duration. A classifier for the prominence of all words occurring in the sentences has to take into account this bias in the training data.
The other application is the disambiguation of two possible interpretations of a sentence, for instance in *uitsluitend VOOR instappen (only get on at front)* versus *uitsluitend voor INstappen (only for getting on)*. In such a disambiguation task there are two words (syllables) involved and it must be decided which of the two words is the most prominent one. To answer this question, the classification has to be based on an unbiased distribution (= equal number of prominent and non-prominent words). Figure 5.2b is an example of this.
To cover both options in the analyses, our training and test data will be split up into a biased and an unbiased set. We are aware of the fact that neural networks are able to account for the prior probability of prominence in the data, but by using the unbiased and biased distributions for training and testing the neural networks are optimally trained for such a task, as mentioned above. So, it makes sense to use biased and unbiased sets as different training conditions.

5.2 Prominence recognition with neural networks

Before going into detail about the contribution of each individual acoustic feature and the performance of an 'optimal' neural network, a brief description of the acoustic features and their pre-processing is given in the following subsections.

5.2.1 Acoustic input features

In chapter 4 the main acoustic features were described in detail and a number of results of this analysis of useful acoustic features for prominence classification were presented. These features are complemented with overall features such as the median F_0 of the sentence. The total set of acoustic features is a set of twelve features as displayed in figure 5.3 and given below.

1. vowel duration;
2. vowel duration normalized for intrinsic vowel duration;
3. sentence speaking rate;
4. vowel intensity normalized for the overall intensity of the given sentence;
5. vowel intensity (sentence normalized) normalized for the intrinsic vowel intensity;
6. overall intensity per sentence;
7. syllable duration;

8. median F_0 per syllable;
9. range of F_0 per syllable;
10. median F_0 corrected for the median F_0 per sentence;
11. median F_0 of the sentence;
12. range F_0 per word.

In figure 5.3 this set of twelve features is displayed as input for a neural network.

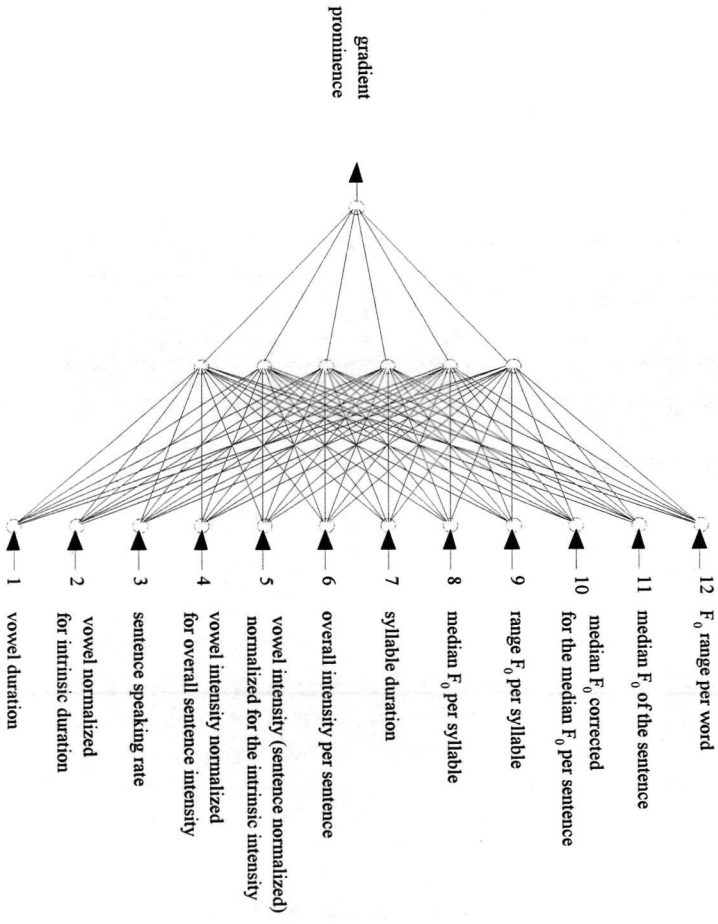


Figure 5.3: An example topology of a neural network, with input features and a single output node, for gradient prominence prediction.

5.2.2 Pre-processing of the input features

The network algorithm implemented in PRAAT prefers inputs between 0 and 1. Therefore all input features are scaled between 0 and 1. This adaptation is a specific requirement for this software and does not apply for neural nets in general. This scaling is performed in the following way: Firstly, per acoustic feature the 99% percentile (max) and the 1% percentile (min) are determined for all training and test data together. (See for more details Appendix A 5.1.) All values occurring beyond these borders are set to 1 and 0, respectively. (Wrong measurements only mildly influence the scaling factor.) The values are scaled by the following formula:

$$x_{scaled} = \begin{cases} 0 & x < \min \\ 1 & x > \max \\ \dfrac{x - \min}{\max - \min} & \text{else} \end{cases}$$

5.2.2.1 Correlation

The set of twelve features obviously contains features with correlated information. For instance, the median F_0 features in corrected (10) and uncorrected (8) form may contain such dependent information. The correlation matrix of all features used is given in table 5.1.

Table 5.1. The correlation matrix of the twelve acoustic features. The acoustic features are indicated with a feature number; the feature description is given in figure 5.3.

Feature number	1	2	3	4	5	6	7	8	9	10	11	12
1	-	0.82	0.23	0.19	0.01	0.12	0.57	0.08	0.26	0.00	0.10	0.29
2	-	-	0.29	0.06	0.05	0.14	0.40	0.11	0.22	0.01	0.14	0.21
3	-	-	-	0.03	0.03	0.24	0.27	0.10	0.12	-0.01	0.15	0.12
4	-	-	-	-	0.90	-0.02	0.28	0.12	0.14	0.18	-0.01	0.17
5	-	-	-	-	-	-0.04	0.09	0.13	0.08	0.20	-0.01	0.09
6	-	-	-	-	-	-	0.08	0.24	0.09	-0.01	0.32	0.09
7	-	-	-	-	-	-	-	0.04	0.37	0.01	0.05	0.34
8	-	-	-	-	-	-	-	-	0.07	0.65	0.71	0.07
9	-	-	-	-	-	-	-	-	-	0.10	0.01	0.78
10	-	-	-	-	-	-	-	-	-	-	-0.06	0.09
11	-	-	-	-	-	-	-	-	-	-	-	0.01
12	-	-	-	-	-	-	-	-	-	-	-	-

The correlations calculated on the original data group into four parts; first high correlations between features that are based on each other such as the vowel duration (1) and the normalized vowel duration (2). Their correlation is 0.82. The second group concerns lower correlations between features that contain similar information such as vowel duration (1) and syllable duration (7). Their correlation is 0.57. The third group shows positive but rather low correlations such as vowel duration (1) and F_0 range per word (12), which correlates 0.29. Fourth are those features that hardly correlate at all, e.g. sentence speaking rate (3) and vowel intensity sentence normalized (4). Their correlation is 0.03.
Some features are included in our analysis simply because we want to test what the effect of various corrections is. It will be tested further on in this chapter whether two highly correlating features should be left out as input features in order to obtain a better performance. As a first step we will train with all twelve input features, neglecting the fact that certain features correlate highly, because in brute force research of, for instance Kießling (1996), the use of as many features as possible leads to high correct classification rates.

5.2.3 Design of the training and testing data

There are several factors that influence the performance of the net. The first is feature representation. Second, the number of hidden nodes in the hidden layer also influences the performance of the net. This number is directly related to the degrees of freedom the neural network has to adjust itself to the training material. Third, the number of iterations will also influence the results. Finally, the distribution over the different prominence categories, as mentioned in section 5.1.2, influences the results. The output needs further specification. One could train and test with discrete output; this is done with separate nodes for each prominence class. Alternatively, one could train and test with continuous output; this requires only one output node. Such a single node with linear output (or with sigmoid) is enough to predict a gradient prominence scale.

The general structure of the test and training set was described in chapter 2, so a summary here will suffice. As described there, the training set consists of 1244 sentences marked for prominence by ten listeners, which resulted in cumulative prominence marks between 0 and 10. By means of a hierarchical cluster analysis the scale was reduced to four classes, namely 0, I, II, III. For simplicity's sake the discrete output was set to two possibilities, namely 'non-prominent' (containing prominence class 0 and I) and 'prominent (containing prominence class II and III). Figure 5.4 gives more details about the distribution of training and test data.
Only one selected 'optimal' listener judged the 1000 sentences of the Independent Test set. With this test set it can be independently tested whether the neural network behaves similarly to one of the listeners. This can only be tested for the binary prominent / non-prominent distinction. We decided to train separate neural networks for prominence degrees i.e. gradient prominence prediction as well as for binary prominence prediction.

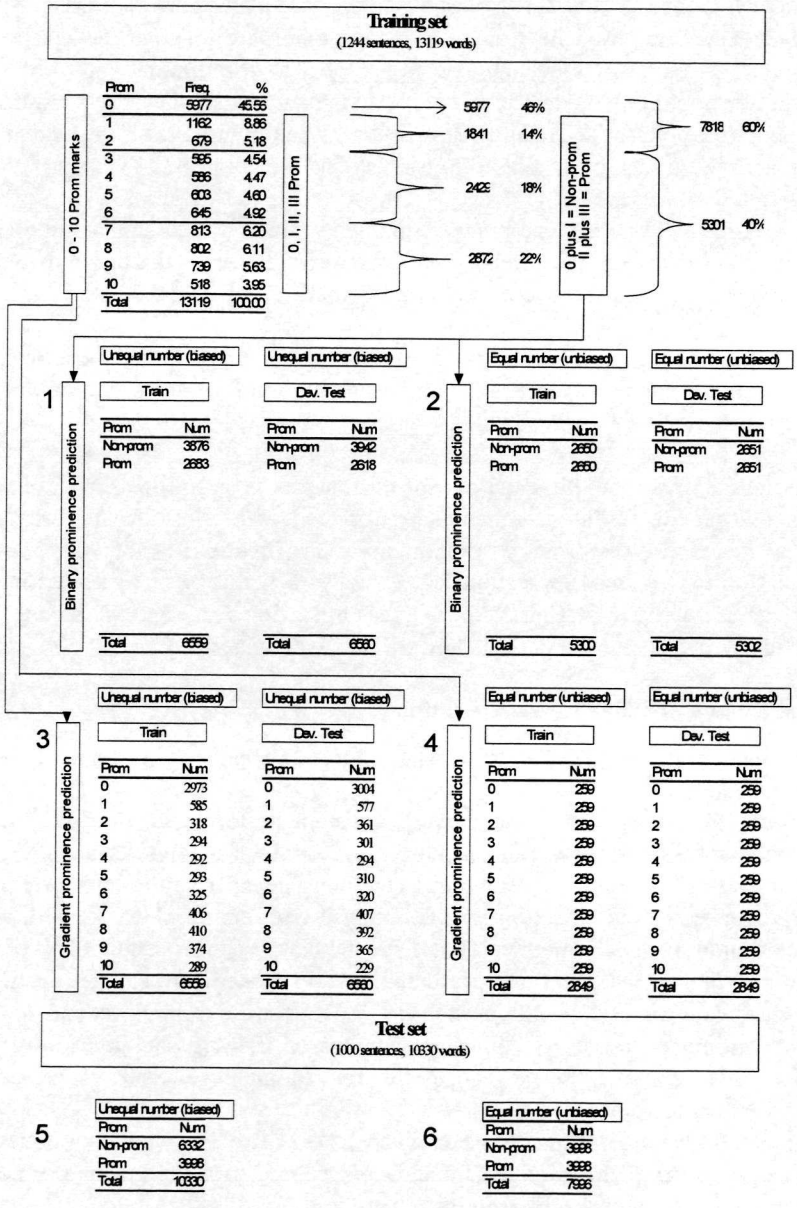


Figure 5.4: Diagram of the Test and Training sets, their distribution over biased and unbiased Training sets and Development test sets, and their distribution over binary and gradient prominence classes. In the Independent Test set only binary categories are used.

In order to train and test the 0-10 gradient prominence scale, the original training set of 1244 sentences with 13119 words, was randomly divided into two equal parts; once for biased and once for unbiased training. See also figure 5.4 dataset 3) and 4). The first part was used for training purposes (Training set) and the second part to test during training (Development test set). The performance of the gradient prediction could only be tested by the Development test set, because a gradient scale was not available for the Independent Test set. (See figure 5.4 for the exact numbers of data in these sets for the biased 5) training and the unbiased 6) version.)

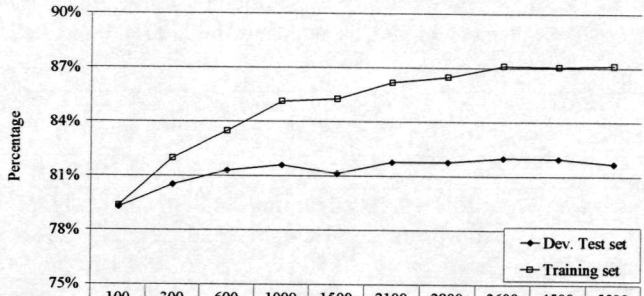
In order to train and test the binary prominence prediction, a similar subdivision of the training data was performed. Our data were randomly divided into two equal parts; one for biased and one for unbiased training. See figure 5.4 dataset 1) and 2).

As for gradient prominence prediction there is also a Training and Development test set. An Independent Test set, one for biased and one for unbiased condition, figure 5.4 dataset 5) and 6), was used to test binary prominence prediction.

The outline of the experimental part of this chapter is as follows. First we will deal with the general results of a neural network fed with all twelve features. Section 5.2.4 describes the binary prominence classification. Gradient prominence prediction is discussed in section 5.2.5. Analyses with a set of selected individual features are given in section 5.2.6, whereas in section 5.2.7 and 5.2.8 combinations of features are the topic. A conclusion will be given in section 5.3.

5.2.4 Binary prominence classification

This section is concerned with dataset 1) and 2) (figure 5.4): binary prominence classification.
In order to get an idea about the classification performance first of all a linear discriminant analysis was run on these data of which the results are given in the upper part of table 5.2. With an LDA only linear relationships are used for classification, so classification results obtained with neural networks with a hidden layer should always be higher. Therefore, the classification results of the LDA are used as a bottom indication. As presented in table 5.2 the correct classification using unbiased data (dataset 2) in figure 5.4) is 77.01% for the training set and 76.05% for the development test set. Training with the biased training data gives lower correct prominence classification, 76.86% for the training set and 75.71% for the development test set.
The neural nets are designed in such a way as to create are two output nodes; one is active when the features of the input vector belong to a prominent word, and the other one is active for a non-prominent word.
Several networks under biased and unbiased conditions were trained with the number of hidden nodes varying from 2 to 18, whereas the number of iterations differed also. Degrees of freedom vary from 32 with a 12-2-2 net up to 272 with a 12-18-2 net. The net decides by the so-called 'winner-takes-all' criterion. Figure 5.5



	100	300	600	1000	1500	2100	2800	3600	4500	5500
Dev. Test set	79.28%	80.52%	81.29%	81.59%	81.14%	81.77%	81.77%	82.01%	81.97%	81.69%
Training set	79.37%	81.98%	83.49%	85.13%	85.27%	86.18%	86.50%	87.11%	87.05%	87.13%

Figure 5.5: Correct recognition percentages of the Training and Development test sets classified with a neural network with the topology of 12-14-2 for the distinction of prominence non-prominence (unbiased training).

presents classification results on the Training and the Development test set with an increasing number of training iterations. This example is taken from table 5.2 and deals with the net with 14 hidden nodes. Whereas the curve for the training data steadily increases, the curve for the development test data starts to decrease somewhere around 3600 training iterations. The number of training iterations needed also depends on the training algorithm. At this turning point the net is sufficiently trained but at the same time is not too closely adjusted to the training data. The results, presented in table 5.2, are selected according to this criterion. Only results of fully trained nets are presented. The recognition rates are presented as overall results and results that have been separated for prominent and non-prominent recognition. Table 5.2 gives the results of several neural nets for the Training set (1, 2 in figure 5.4) and for the Development test set (1, 2 in figure 5.4). The percentages correct prominence classification with neural networks are always higher than classification with an LDA. The correct classification on the training data also are always higher than on the development test set data. Neural networks with more than 6 hidden nodes in their hidden layer give better performances than with less hidden nodes (see table 5.2).

Unbiased training (equal numbers) achieves better recognition results for prominence classification (around 82%) than for non-prominence classification (around 77%). The features describing the prominent words may be more clearly defined than the features describing non-prominent words. The opposite is true for the biased trained neural networks. There the non-prominence classification (around 82%) achieves higher recognition results than the prominence classification (around 76%). This is explained by the fact that a neural net trained with unequal numbers is biased to recognize the more frequently occurring non-prominent words. This

Table 5.2: Results (percent correct classification) of several neural networks with a variable number of hidden nodes in the hidden layer, trained with biased und unbiased data.

		Unbiased Training			Biased Training		
		Non-prom	Prom	%	Non-prom	Prom	%
Linear discriminant analysis (LDA)	All	Training set		77.01	Training set		76.86
	Non-prom	1996	654	75.32	2986	956	75.75
	Prom	564	2086	78.71	562	2056	78.53
	All	Dev. test set		76.05	Dev. test set		75.71
	Non-prom	1960	691	73.93	0.74355	994	74.87
	Prom	579	2072	78.16	0.77674	2084	78.79
Num hidden nodes							
2	All	Training set		**79.79**	Training set		**78.82**
	Non-prom	1993	657	75.21	3135	741	80.88
	Prom	414	2236	84.38	648	2035	75.85
	All	Dev. test set		**78.06**	Dev. test set		**78.09**
	Non-prom	1977	674	74.58	3161	781	80.19
	Prom	489	2162	81.55	656	1962	74.94
4	All	Training set		**81.13**	Training set		**80.99**
	Non-prom	2056	594	77.58	3208	668	82.77
	Prom	406	2244	84.68	579	2104	78.42
	All	Dev. test set		**79.80**	Dev. test set		**79.77**
	Non-prom	2042	609	77.03	3238	704	82.14
	Prom	462	2189	82.57	623	1995	76.20
8	All	Training set		**82.49**	Training set		**81.06**
	Non-prom	2083	567	78.60	3218	658	83.02
	Prom	361	2289	86.38	584	2099	78.23
	All	Dev. test set		**80.33**	Dev. test set		**79.85**
	Non-prom	2050	601	77.33	3234	708	82.04
	Prom	442	2209	83.33	614	2004	76.55
10	All	Training set		**84.26**	Training set		**83.34**
	Non-prom	2157	493	81.70	3279	597	84.60
	Prom	341	2309	87.49	496	2187	81.51
	All	Dev. test set		**81.25**	Dev. test set		[illegible]
	Non-prom	2059	592	77.79	3228	714	81.89
	Prom	402	2249	85.08	591	2027	77.43
14	All	Training set		**86.66**	Training set		**80.81**
	Non-prom	2232	418	84.67	3227	649	83.26
	Prom	289	2361	89.54	610	2073	77.26
	All	Dev. test set		[illegible]	Dev. test set		**79.91**
	Non-prom	2114	537	79.73	3254	688	82.55
	Prom	417	2234	84.28	630	1988	75.94
18	All	Training set		**82.55**	Training set		**81.26**
	Non-prom	2106	544	79.47	3247	629	83.77
	Prom	381	2269	85.62	600	2083	77.64
	All	Dev. test set		**80.20**	Dev. test set		**79.54**
	Non-prom	2068	583	78.01	3256	686	82.60
	Prom	467	2184	82.38	656	1962	74.94

shows that the distributions of prominent and non-prominent words play a role in the classification process.

In the unbiased case the best overall performance of prominence classification is 82.01% on the Development test set. This result was achieved in a neural network with 14 nodes in the hidden layer. The best performance for a biased trained network (10 hidden nodes) resulted in an overall recognition rate of 80.11%.

5.2.4.1 Testing with the Independent test set

The best nets based on the general results of table 5.2 were selected for independent testing. For the biased condition this is net 12-10-2, and for the unbiased condition it is net 12-14-2. These 'optimal' neural networks were used to mark the words in the 1000 sentences of the Independent Test set (figure 5.4 dataset 5 and 6). These prominence-marking results can easily be compared to the marks of the one 'optimal' listener who marked these words for prominence. The results are presented in table 5.3 and 5.4 in the form of a crosstable. Table 5.3 concerns the condition of the unequal numbers of prominent and non-prominent words (biased data) in the 1000 sentences of the Test set. The input data in table 5.4 deal with the unbiased case. The number of non-prominent words is then randomly reduced to the same number of available prominent words, which is 3998 (see condition 6 in figure 5.4).

Table 5.3: The correct recognition rates of prominence classification of the Independent Test set (biased data, condition 5 in figure 5.4). The total numbers as well as the percentages are given for the networks trained under unbiased and biased condition; the networks with the topology of 12-14-2 and 12-10-2 were optimal.

	Unbiased training			Biased training		
	Non-prom	Prom	%	Non-prom	Prom	%
	Test set (biased)		77.1	Test set (biased)		78.9
Non-prom	4907	1425	77.5	5232	1100	82.6
Prom	942	3056	76.4	1079	2919	73.0
Measure of agreement (κ)			0.53			0.57

First, we give a description of table 5.3. Percentages of prominence and non-prominence recognition rates are 77.5% and 76.4%, respectively, for the unbiased training condition (on average 77.1% correct). Non-prominent recognition is 82.6% for the biased training condition. This is comparable to the results of the development test set. Training with biased data gives better results on non-prominent recognition; the prominent recognition drops to 73.0%. A biased trained net performs best (78.9%) on the Independent Test set. The better performance of the biased trained net is as expected, as the distribution of the prominence marks in the 1000 sentences of the Independent Test set coincides with the distribution in the

training material. The Test set contains 10330 words; 3988 of these are marked as prominent, which is 39%. The remaining 61% is non-prominent. The 1244 sentences used for the training and development test set contain 13119 words, of which 7818 are treated as prominent, which is 40% of the total number of words.

When the bias in the data of the Test set is removed, i.e., when equal numbers of prominent and non-prominent words are presented, the performance of the net 12-10-2 decreases to 77.78% correct classification (see table 5.4). The performance of the unbiased trained net 12-14-2, tested with unbiased data of the Test set (condition 6 in figure 5.4) is 77.05% correct. Training and testing under the same biased or unbiased conditions do not give consistently better classification results. In table 5.3 the results are better when trained and tested with biased data, but contrary to this in table 5.4 the results for training and testing with unbiased data shows less percentages correct than trained with biased data and tested with unbiased data.
The between-listener agreement (section 2.4.1.2) expressed in Cohen's Kappa κ, was calculated for the results of the neural network and the listener who marked all 1000 test sentences. Kappa values are 0.53 (unbiased training) and 0.57, (biased training), see table 5.3. In an unbiased Test set (table 5.4) these Kappa values hardly differ. Similar values (on average $\kappa = 0.50$; Std. Dev. = 0.16) were measured for the between-listener agreements see section 2.4.1.2. This means that the neural network behaves similarly to any listener, and that the differences in prominence classification are as accurate as the prominence classification of any naive listener. The performance of the net is indistinguishable from any listener.

Table 5.4: This table presents the recognition rates of prominence classification on the Independent Test set (unbiased data, in figure 5.4 condition 6). The total numbers as well as the percentages are given for the 'optimal neural networks' trained under unbiased and biased condition with the topology of 12-14-2 and 12-10-2, respectively.

	Unbiased training			Biased training		
	Non-prom	Prom	%	Non-prom	Prom	%
	Test set (unbiased)		77.05	Test set (unbiased)		77.78
Non-prom	3105	893	77.66	3300	698	82.54
Prom	942	3056	76.44	1079	2919	73.01
Measure of agreement (κ)			0.54			0.56

5.2.4.2 Summary and conclusion

For the binary prominence classification the following results have been reached: 82% correct classification on a Development test and 79% correct on an Independent Test set. The performance may be accurate enough to allow sentence disambiguation to be done by such a classifier, especially if one keeps in mind that

any listener is indistinguishable from our 'optimal' neural network (79% correct, κ = 0.57, table 5.3) fed with acoustical information only. Considering that the listener has both acoustical and linguistic information, a combination of both acoustic and linguistic input features may further improve the classification.
Comparing correct recognition rates of an LDA with neural networks shows that apparently non-linear relationships exists between acoustic input features and prominence.
With an LDA only linear relationships can be used for prominence prediction, however, with neural networks with a hidden layer higher order relationships can be exploited. The prominence classification with a neural network appears to be always better than the classification with an LDA.

5.2.5 Gradient prominence prediction

In the previous section we discussed binary prominence classification only. Training a neural network that provides gradient prominence, e.g. for our approach a linear output of the neural net, has the advantage that it is not limited to two prominence classes. This gradient output can always be reduced to two or more discrete prominence classes. This facilitates comparison between binary and gradient classification. Another advantage is that only one output node is needed. This reduces the degrees of freedom substantially while the prediction of any amount of prominence is still possible. Because of the relatively low number of degrees of freedom, there is no problem in having insufficient training material, which lowers the danger of overtraining. This section deals with data selection 3 and 4 in figure 5.4. This time, however, we use a prominence scale from 0-10.
Several neural networks were trained under different conditions. However, all twelve features are always used as input. The number of hidden nodes varies from 2 to 18. Only nets with one hidden layer were used. The output of such nets is a single value around 0. Therefore the original cumulative prominence marks of the listeners (0-10) were scaled between -1 and +1 by using the formula 1/5 * prom -1. For instance prominence mark 9 yields 0.8 and prominence mark 4 yields -0.2.
As described above, the neural networks were trained with various numbers of training iterations. The number of iterations ranges from 100 up to 5500. The Training and Development test sets are biased (3 in figure 5.4) or unbiased (4 in figure 5.4). In order to present an overall performance of all the trained neural networks with varying number of training iterations, the linear correlations between the predicted prominence and the perceived prominence are calculated for the Training and the Development test set. Such a relationship may not be linear, but a higher order correlation was not tested. These linear correlation coefficients indicate the performance of the trained neural network and are used to select the optimal network. The results of neural networks giving a linear output are difficult to present. Each input feature has an output of around 0, for instance 0.4563. For rescaling the output we use the formula ((output + 1) * 5). This gives us the predicted prominence value of 7.2815 (on the original scale from 0 to 10).

5.2.5.1 Results

For the optimal unbiased trained network the highest correlation coefficient is 0.60 for the Development Test set (unbiased) and 0.64 for the Training set. This net has a topology of 12-6-1. For the unbiased condition the highest correlation coefficients are 0.70 (Development Test set) and 0.72 (Training set). The optimal net with the topology of 12-10-1 achieved these highest correlation coefficients.
In order to graphically compare the predicted prominence with the perceived prominence, figures 5.6 and 5.7 give medians (of the predicted prominence within

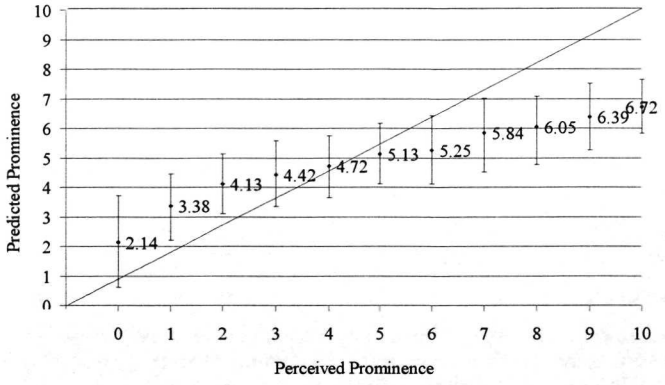


Figure 5.6: Medians ± 1 IQR (Inter Quartile Range) per median of the predicted prominence (trained with unbiased data) of the Development test set on the perceived prominence scale of 0-10. The linear correlation is r = 0.60. The dashed line gives the perfect prediction.

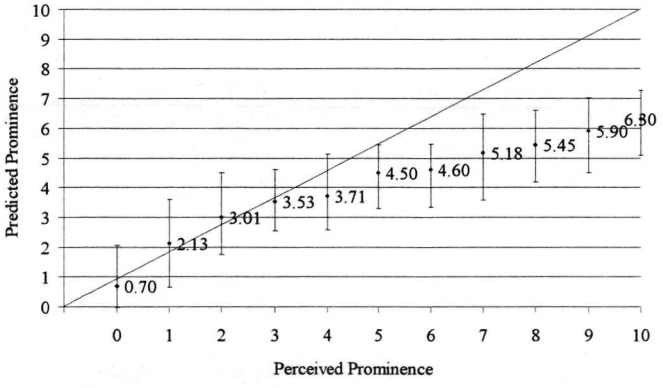


Figure 5.7: Median ± 1 IQR of the predicted prominence (trained with biased data) of the Development test on the perceived prominence scale of 0-10. The linear correlation is r = 0.70.

the prominence scale of 0-10) for the optimal neural networks for the Development test set. As error bars the Inter Quartile Ranges per median (75% and the 25% percentiles) are plotted.
The extremes of the perceived prominence scale are shifted towards the middle of the predicted prominence scale. The calculated percentiles confirm this: perceived prominence of 8, 9 and 10 is predicted with median values of 6.05, 6.39 and 6.72, respectively (figure 5.6). These values are even further shifted to the mid range for the biased training, namely median values of 5.45, 5.90 and 6.30, respectively (figure 5.7). However, for the biased training the predicted median value of 0.70 is closer to the perceived prominence 0. This is not the case for the unbiased training, where the perceived prominence of 0 is predicted to be 2.14. The error bars, expressed in IQR, show that there is a lot of overlap, indicating that the classification of each individual prominence value of 0-10 is not an easy task. In order to examine how well the predicted gradient prominence values fit into the perceived prominence scale (0-10) confusion matrices are constructed.

Tables 5.5 and 5.6 present confusion matrices of the perceived prominence versus the predicted prominence scales, achieved by the optimal neural networks (12-6-1 and 12-10-1) respectively, which have the highest correlation coefficients. In order

Table 5.5: Gradient prominence prediction versus perceived prominence. This matrix is based on the Development Test set. The prediction was achieved with unbiased data and the topology of the net was 12-8-1. Cells containing > 30 data points are boxed.

Predicted prominence \ Perceived prominence	0	1	2	3	4	5	6	7	8	9	10	Total
-2	-	-	-	-	-	-	-	-	1	-	-	1
-1	20	4	-	-	-	-	-	-	1	-	-	25
0	37	10	2	1	-	-	-	-	-	-	-	50
1	43	28	12	5	7	2	-	-	-	1	-	98
2	46	36	22	23	18	11	6	5	5	1	1	174
3	36	61	51	46	33	27	26	11	12	7	1	311
4	40	59	69	61	65	51	60	47	32	27	16	527
5	25	33	51	56	50	61	54	46	46	36	34	492
6	9	16	27	38	54	62	51	56	61	65	58	497
7	2	9	20	20	23	34	41	48	59	57	77	390
8	1	3	5	8	7	7	19	39	34	42	42	207
9	-	-	-	1	1	4	2	7	6	20	23	64
10	-	-	-	-	1	-	-	-	2	3	7	13
Total	259	259	259	259	259	259	259	259	259	259	259	2849

to put the predicted amount of prominence in such a confusion matrix, the linear output is placed on a scale of 0-10, as explained above, by rounding these values to integers. For instance the number 7.2815 is rounded off to 7. This number yields one tally in row 7 of the predicted prominence.

Especially the cells in the middle of the matrices cause confusion. The perceptually very prominent categories (8, 9, 10) are often predicted with less prominence. Training with equal numbers provides a poor non-prominence prediction; perceived prominence 0 and 1 is often predicted with 2 and 3 (see table 5.5). In addition, it is also possible that predictions are less than 0 and greater than 1. This depends exclusively on the input values. It can occur that if the neural net is fed with deviant data (for instance, because of a measurement error) the output of the neural net is also very deviant. In fact two words with a perceived prominence of 8 are predicted with a value less than 0, namely -1 and even -2. This holds for the unbiased condition. This is obviously wrong. We decided to keep these values, because it concerns only a few incidental cases.

For the training with unequal numbers (biased) the prediction on the lower part of the prominence scale is much better than for unbiased training; the cells around 0 are more filled.

Table 5.6: Gradient prominence prediction versus perceived prominence. This matrix is based on the Development Test set. The prediction was achieved with biased data and the topology of the neural network was 12-10-1. Cells containing > 10% of the total per perceived prominence scale are boxed in.

Predicted prominence \ Perceived prominence	0	1	2	3	4	5	6	7	8	9	10	Total
-5	1	-	-	-	-	-	-	-	-	-	-	1
-3	3	-	-	-	-	-	-	-	-	-	-	3
-2	20	-	-	1	1	-	-	-	-	-	-	22
-1	218	23	2	-	-	-	-	2	1	-	-	246
0	1108	106	13	12	6	8	3	1	2	-	1	1260
1	651	100	55	29	22	6	13	10	5	2	1	894
2	404	102	79	32	42	29	31	33	20	9	5	786
3	271	94	64	73	62	46	49	52	36	25	9	781
4	175	75	61	70	55	66	60	62	55	57	18	754
5	104	52	47	36	53	84	86	72	84	63	40	721
6	34	14	28	23	33	40	36	81	80	69	56	494
7	13	9	11	18	13	25	31	55	85	79	51	390
8	2	2	1	7	7	2	8	30	19	42	35	155
9	-	-	-	-	-	4	3	9	4	18	13	51
10	-	-	-	-	-	-	-	-	1	1	-	2
Total	3004	577	361	301	294	310	320	407	392	365	229	6560

5.2.6 Analyses of individual features

In chapter 4 several individual features were analyzed in general terms as prominence predictors and several histograms concerning these acoustic features were given. These histograms indicate the ability to distinguish between the two extremes of the prominence classes (namely prominence categories 0 and III). The discriminative ability of these features will be analyzed in more detail in this subsection by using neural net techniques and will be compared also with the relevant histograms as displayed in chapter 4. All four prominence categories are involved in the analyses presented in this chapter (0 and I as non-prominent, and II and III as prominent). In chapter 4 we restricted ourselves to the two extremes of the prominence categories (0 and III). Techniques such as linear discriminant analyses or CART-trees may be more powerful, but in this study simple feed-forward networks do help us to investigate acoustic correlates in detail. Our simple 1-2 neural networks give the similar results as an LDA. We chose to stay with neural networks. Individual networks with one input node and two output nodes are trained with eight of the twelve features. Features giving overall information of the sentence are omitted. Eight of the twelve features, as described in figure 5.3, were used as such single features; namely the vowel duration (1), vowel duration normalized for intrinsic vowel duration (2), vowel intensity normalized for the overall intensity of the given sentence (5), vowel intensity (sentence normalized) normalized for the intrinsic vowel intensity (6), syllable duration (7), median F_0 corrected for the median F_0 per sentence (9), range of F_0 per syllable (10), range F_0 per word (12).
Later on in this chapter we will also study some combinations of features. Individual features were analyzed by training a simple neural network with one input node and two output nodes (prominent or non-prominent). Such simple networks can be used to analyze the individual input features in two different ways. On the one hand the performance of the classification with single acoustic input features can give information about the discriminative power of these individual features, and on the other hand the neural networks themselves can be analyzed. If one uses neural networks without a hidden layer, only linear relations can be estimated, but for a preliminary examination of the data this will suffice.
As said above the design of the Training and Development Test sets are available in a biased and in an unbiased version. Consequently, a total of 16 neural networks is required that will vary only in one input feature. However, the difference between biased and unbiased training and testing remains. These neural networks were mostly trained with only 18 training iterations, which is sufficient given the few degrees of freedom. These simple nets have only 4 variables to adjust to the data.
Before presenting the performance of these 16 neural networks, the network itself will be analysed.

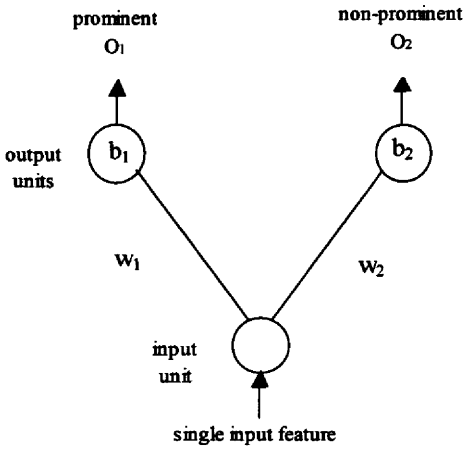


Figure 5.8: A neural network with the topology of one input node and two output nodes (1:2). Such networks are used to analyze single features and to estimate critical boundaries.

5.2.6.1 Analyzing the neural network

The neural nets used are designed in such a way that the first output node fires when the input concerns a prominent word and the second output node fires when the input concerns a non-prominent word. It can be calculated where the trained neural network puts the decision threshold in the training session. Beyond this threshold the net decides to classify this word as prominent and below this threshold value the net marks this word as being non-prominent.

The resulting activation threshold for our minimal neural network can be expressed as follows, where O_1 and O_2 are the output functions of the two output units (see figure 5.8):

$$O_1 = \frac{1}{1+e^{-(+w_1 \cdot Input - b_1)}}$$

$$O_2 = \frac{1}{1+e^{-(+w_2 \cdot Input - b_2)}}$$

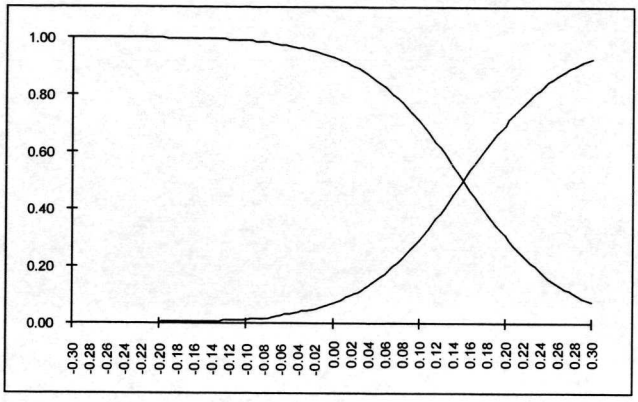


Figure 5.9: The two activation functions for the output nodes. There is one crossover point near 0.15.

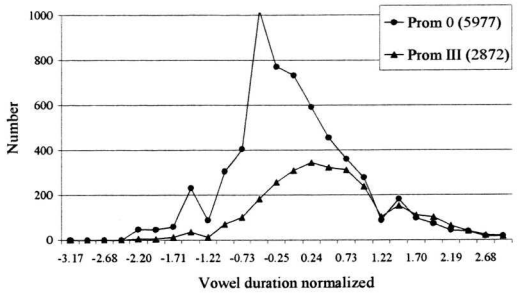
The crossover point is calculated by equalizing these two functions. This crossover point is displayed in figure 5.9. This is exactly the critical point; at this point a prominent and a non-prominent word are equally probable.

The values of these crossover points are estimated for the eight individual features. Table 5.7 gives the values for the biased and for the unbiased versions. By comparing the two columns it can be seen that the thresholds estimated by using a neural network trained with biased training input (1 in figure 5.4) are shifted to longer vowels and syllables. These vowels have also a higher intensity and show greater ranges in their F_0 movements than in the unbiased version (2 in figure 5.4). For 'vowel duration' the critical value is 0.10 s or 0.12 s, respectively. These values

Table 5.7: The estimated boundaries for 8 acoustic features. These boundaries function as a crossover point; beyond these values the neural network classifies the given data as belonging to a prominent word.

Feature	Estimated boundaries	
	Unbiased	Biased
Vowel duration (s)	0.10	0.12
Vowel duration normalized (z-score)	0.09	0.75
Vowel normalized for sentences intensity (dB)	77.89	79.69
Vowel intensity (dB) normalized (z-score)	0.17	0.74
Syllable duration (s)	0.21	0.24
Range F_0 per syllable (st)	3.36	4.19
Median F_0 per syllable (st above the sentence median F_0)	0.23	4.33
Range F_0 per word (st)	3.75	4.69

a) Biased:



b) Unbiased:

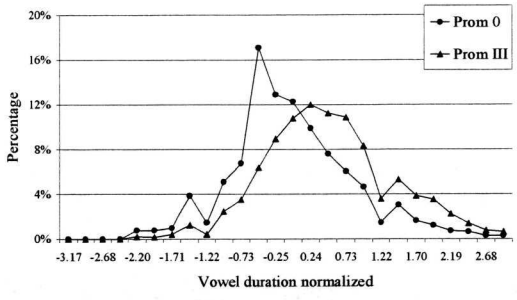


Figure 5.10: The biased and unbiased distribution of the vowel duration normalized for intrinsic vowel duration. The peaks at the edges of the histograms are due to the bin distribution. (More details are explained in section 4.2.2.3.)

indicate that vowels longer than 0.10 s or 0.12 s are classified as prominent. This unbiased threshold of 0.10 s is similar to the crossover point in figure 4.9, which actually represents unbiased data, because in these figures percentages are presented. The threshold does not have to be identical because the data are not identical (in the histogram only prominence class 0 and III are presented, whereas the nets are trained with 0 and I as 'non-prominent' and II and III as 'prominent'). The unbiased threshold (0.09 s) of 'vowel duration normalized for intrinsic vowel duration' shows that, if a vowel is classified as prominent (prominence class II and III) it must be longer than the threshold value of the class it belongs to. If the normalized duration had been exactly equal to the mean of its class, this value would have been 0. The network trained with biased input data of the vowel duration corrected for the intrinsic vowel duration (vowel duration normalized) places the threshold at 0.75 (z-score). This is a large shift to the right. This can be explained by looking at the distribution of this feature as shown in figure 5.10. Graph a) presents the biased data of the two extremes of the prominent class (0 and III), and graph b) presents percentages, which are corrected for the differences in numbers in the two extremes of the two prominence classes. Prominence class III lies below the prominence class 0 histogram as shown in graph a). The neural network places the threshold far to the

right so that chance distribution of the frequency of occurrence determines the classification.
A similar situation occurs for 'vowel intensity corrected for intrinsic vowel intensity' and for 'median F_0 per syllable'. These three features (vowel duration normalized for intrinsic vowel duration, vowel intensity (dB) normalized for intrinsic vowel intensity, and the median F_0 per syllable) are probably not useful in a biased classification situation. The classification results, presented later in this section, will show whether these three features are useful for unbiased classification situations, e.g. for disambiguating the meaning of sentences. The thresholds of the other features related to the vowel intensity are as expected. The 'biased' thresholds have further shifted towards greater intensity in order to cope with the greater probability that data belong to non-prominent words (60% of the data belong to non-prominent words, whereas only 40% belong to prominent words). The F_0 range thresholds measured on syllables (3.36 st or 4.19 st) and on words (3.75 st or 4.69 st), respectively, show that the F_0 range must be larger for the word condition than for the syllable condition. The thresholds belonging to the unbiased version correspond to the crossover points of the histograms as in figure 4.6.

5.2.6.2 Analyzing the performance of the individual features

The performances of these simple neural networks are of interest because a comparison can be made with the amount of overlap of the histograms as displayed in chapter 4 and as expressed in table 4.3 in that chapter. In this comparison, not only the absolute recognition rate is important but also whether a ranking can be made in order to work out the most useful features for prominence classification. A combination of the 'best' features brings us closer to the training of the 'optimal' neural network for ultimate prominence classification for a binary prominent / non-prominent distinction as well as a gradient prominence classification.
Table 5.8 presents the recognition rates for the above-discussed simple neural networks with single input features for the biased and unbiased versions of training and testing. The recognition rates are only given for the Development Test set, and are separately presented for prominent and non-prominent input. The chance level of the frequency of occurrence is 50% for the unbiased recognition rates, whereas for the biased version this level is about 60% according to the prominent / non-prominent distribution as explained above. As expected from the estimation of the thresholds, 'vowel duration normalized for intrinsic duration', and 'vowel intensity normalized for the intrinsic intensity', and 'median F_0 per syllable', performed slightly above chance level for the biased trained neural networks. Because of the differences in the biased distribution, the non-prominent recognition in biased condition is always better than the prominent recognition. In the case of the unbiased condition, recognizing prominent words is better when using intensity features; the other features do not show such a constant preference.

Table 5.8: The performance of the individual acoustic features trained with a simple neural network (1-2) without a hidden layer. The overall recognition rate on the Development test set for the biased and the unbiased data is given together with the separate results for the prominent and non-prominent recognition. Also the ranking of the overall scores is given together with the ranking based on the histogram overlap as presented in chapter 4, section 4.3, table 4.3.

		Unbiased	Ranking	Biased	Ranking	Ranking overlap
Vowel duration	All	**65.97%**	5	**66.01%**	4	6
	Non-prom	65.87%		81.05%		
	Prom	66.07%		43.88%		
Vowel duration normalized	All	**61.01%**	8	**61.57%**	7	8
	Non-prom	62.23%		83.33%		
	Prom	59.78%		29.57%		
Vowel intensity corrected overall	All	**66.86%**	4	**65.67%**	5	4
	Non-prom	59.86%		74.16%		
	Prom	73.87%		53.18%		
Vowel intensity normalized	All	**62.11%**	7	**61.63%**	6	7
	Non-prom	56.69%		77.95%		
	Prom	67.54%		37.63%		
Syllable duration	All	**71.34%**	2	**70.14%**	1	3
	Non-prom	71.18%		79.82%		
	Prom	71.50%		55.89%		
Range F_0 syllable	All	**69.02%**	3	**67.73%**	3	2
	Non-prom	71.38%		77.18%		
	Prom	66.67%		53.82%		
Median F_0 corrected	All	**64.45%**	6	**60.11%**	7	5
	Non-prom	65.95%		90.14%		
	Prom	62.95%		15.93%		
Range F_0 word	All	**71.85%**	1	**69.83%**	2	2
	Non-prom	72.80%		77.26%		
	Prom	70.90%		58.91%		

Similar to the ranking based on the overlap of the histograms in table 4.3, a discriminability ranking can also be derived from the recognition rates of the neural networks. In table 5.8 such a ranking is displayed for the biased and the unbiased condition. Rankings 2 and 1 are interchanged for the unbiased and biased condition, indicating that syllable duration and F_0 range per word are the features with the highest discriminatory power in both conditions. The F_0 range per syllable follows

these two features. Places 4 and 5 are interchanged as well; this means that the unnormalized version of vowel duration and intensity (sentence normalized) are interchanged. The ranking for the histogram overlap (see section 4.3 table 4.3) follows more or less the network ranking. Place 2 is interchanged with place 3. As already indicated in chapter 4, duration and intensity normalization does not further improve prominence classification. The percentages of correct recognition are poorer for normalized features than for the unnormalized counterparts (see table 5.8).

5.2.6.3 Summary and conclusion

The ranking of the recognition results differs somewhat for the biased and unbiased version, as well as in comparison to the ranking of the amount of overlap (see chapter 4), but the trend is clear. F_0 ranges per word and syllable duration are useful features. Vowel intensity and vowel duration are also useful, while their normalized counterparts do not improve the classification of this speech material.
In the next section it will be estimated whether certain combinations of features will further improve the classification task.

5.2.7 Analyzing combinations of features

Only a few paired feature combinations were selected to train a neural network and, for simplicity reasons, only the results of the unbiased trained nets are presented here. Preferably, only combinations should be involved which do not intercorrelate highly (see table 5.1), ensuring that each feature with the combination contributes independent information.
The recognition results of the various combinations are presented in table 5.9. The

Table 5.9: The correct recognition rates (% correct) of a number of acoustic feature combinations, expressed in overall correct recognition rates on the Development Test set for the unbiased data.

Feature combination	Dev. Test (%)
Range F_0 word - syllable duration	75.50
Range F_0 word - vowel intensity	74.48
Range F_0 word - vowel duration	74.04
Syllable duration - vowel intensity	72.69
Syllable duration - sentence speaking rate	71.45
Vowel duration - vowel intensity	69.33
Vowel intensity - overall intensity	65.98
Vowel duration - sentence speaking rate	66.01
Vowel intensity normalized - overall intensity	62.77
Vowel duration normalized - sentence speaking rate	61.15

results are ordered according to recognition rates. It is striking that the range of F_0 per word, in combination with the syllable duration, vowel intensity, or vowel duration gives the best classification. The acoustic feature combination of syllable duration and vowel intensity also gives acceptable results.
Although the sentence speaking rate yields some effect (chapter 4, figure 4.13), this additional information did not provide an improved performance. The feature of sentence speaking rate in combination with either vowel duration or syllable information gives a performance of 66.01% or 71.45%, respectively. However, the single features analyses gives 65.97% and 71.34% performance. In addition, the combination of sentence speaking rate and vowel duration normalized for intrinsic vowel duration does not improve performance (61.15% with sentence speaking rate versus 61.01% for the single feature). The trained networks with two-wise feature combinations indicate that the unnormalized features are the most promising ones. For this kind of speech material, normalizations do not increase the prominence classification substantially. Normalization cannot (yet) be implemented in such a way that it increases the performance.

5.2.8 Prominence classification with an 'optimal' feature combination

Based on these results we combined F_0 range per word, syllable duration, vowel duration and vowel intensity as a promising feature set.
A neural network trained with these four 'basic' features performs almost as well as

Table 5.10: Performance on the Development Test set of several neural networks with 4 'basic' acoustic input features.

		Unbiased training			Biased training		
Num of hidden nodes		Non-prom	Prom	%	Non-prom	Prom	%
4	All	Dev. test		**78.11**	Dev. test		**78.02**
	Non-prom	1871	655	74.07	3145	760	80.54
	Prom	451	2075	82.15	682	1973	74.31
6	All	Dev. test		**77.95**	Dev. test		[illegible]
	Non-prom	1863	663	73.75	3163	742	81.00
	Prom	451	2075	82.15	682	1973	74.31
8	All	Dev. test		**78.35**	Dev. test		**78.26**
	Non-prom	1871	655	74.07	3169	736	81.15
	Prom	439	2087	82.62	690	1965	74.01
10	All	Dev. test		**78.11**	Dev. test		**77.94**
	Non-prom	1874	652	74.19	3128	777	80.10
	Prom	454	2072	82.03	670	1985	74.76
12	All	Dev. test		**78.48**	Dev. test		**78.08**
	Non-prom	1869	657	73.99	3150	755	80.67
	Prom	430	2096	82.98	683	1972	74.27
14	All	Dev. test		[illegible]	Dev. test		**77.96**
	Non-prom	1900	626	75.22	3153	752	80.74
	Prom	433	2093	82.86	694	1961	73.86

a neural network trained with all twelve features. The neural network trained with the four features shows an overall performance of 79.04% correct prominence and non-prominence classification on the Development test set under the unbiased condition (see table 5.10). The best biased condition gives a performance of 78.29% correct. The performance of a net trained with all twelve features still performs somewhat better though (unbiased condition 82.01%, biased condition 80.11%, see table 5.2) on the Development test set.

5.3 Discussion and conclusion

Twelve acoustic input features for binary prominence prediction yield 82% (unbiased condition) and 80% correct (biased condition) classification on the Development test set and 79% and 78% correct classification on the Independent test set. Higher results are found in the literature. Kießling (1996) achieved 82.8% (spontaneous speech material) and 95% (read aloud speech material, simple sentences) correct classification (for accent / non-accent in their terminology), using a net with 276 input features including textual information. Such a comparison is not completely fair, because we aim at different goals. The statistical or brute force method used in Kießling (1996) aimed at high recognition rates whereas our approach aimed different perspectives of acoustics and classification. Our prominence classification results in this chapter were achieved using solely twelve acoustic input features.

The binary prominence predictions are as consistent as listeners are. On an Independent Test set (unbiased data) these nets achieve 77% correct ($\kappa = 0.56$) with the most optimal listener (see also chapter 2). The agreement is even higher for the biased data: namely $\kappa = 0.57$. These agreements do not differ from the agreement between listeners which is on average $\kappa = 0.50$. Thus our neural networks are indistinguishable from naive listeners for assigning prominence.

The attempt to predict gradient prominence is much more complicated than binary prominence prediction. A correlation of $r = 0.60$ (unbiased condition) and $r = 0.70$ (biased condition) is achieved. In principle high correlations could only be a indication of high recognition results. Looking at the underlying confusion matrices it appears that the middle range of the prominence scale (0-10) is an area of confusion, and that the extreme of 10 (very prominent) is rarely predicted. It may be that our design of perceptual prominence judgments is not constructed to allow for a really accurate prediction of a gradient scale.

The analyses of the individual acoustic features confirm that the four 'basic' acoustic features, namely vowel duration and intensity, syllable duration and F_0 range per word, yield performances of 79% (unbiased condition) and 78% (biased condition) correct prominence classification on the Development test set. Despite the fact that normalizations provide no further improvement when used as two-wise features, using all twelve input features still shows somewhat better performance

(table 5.2). Exactly what the effect is of the eight remaining features still needs detailed investigation. Also, it can be concluded that the use of a hidden layer provides more accurate prominence classification (table 5.2), also in comparison with an LDA. This means that there is no simple linear relationship between prominence and acoustic features. Therefore, the linear representation in the histograms of chapter 4 is a first approximation only.

6

GENERAL CONCLUSION AND DISCUSSION

Abstract

In this final chapter the various threads in the previous chapters are put together. We demonstrated that naive listeners are able to assign in a consistent way word prominence in individual Polyphone sentences, with little instruction to the listener. This enables prominence annotation for large speech databases by non-experts. Most differences between-listeners (reliability) and within-listeners (consistency) can be ascribed to level shift or level differences. These findings allow us to use an operational definition of prominence. Next, we have demonstrated that, on the basis of text input only, the prominence level of words can be predicted with a performance similar to that of naive listeners. The linguistic information used was limited to word class (POS), word length, Adjective-Noun combinations and the first content word in a sentence. We studied a selected set of acoustic correlates of prominence, all based on F_0, duration and intensity extracted at units not larger than a word. Finally, we used twelve acoustic features as input for a neural net classifier. A binary prominence classifier appeared to be correct in 79% of the cases, which is statistically indistinguishable from a consistent naive labeler. We can thus conclude that automatic prominence assignment is sufficiently powerful to simulate naive labelers for read-aloud declarative medium length sentences in Dutch.

6.1 Introduction

The perceptual concept of prominence, which is the amount of emphasis put on syllables and words that make them 'stand out' in their environment, can function as an interface between acoustics and linguistics. It is strongly related to information structure in terms of 'given' and 'new' information or in terms of focus. In our approach a perceptually defined concept of prominence is the basis for all further analyses. This perceptual phenomenon of prominence appears to be intuitively clear to non-experts. A proper modeling of prominence, either from the signal, or from text, may be useful for several applications in speech technology (e.g. in dialogue handling).
The purpose of this study was to explore various aspects of prominence. First, we developed a useful operational definition of prominence via of judgments of naive (native) listeners. We focused on analyzing various linguistic and acoustic correlates of prominence referring to bottom-up and top-down information, respectively. The final goal of this study was to find and extract those features that can best be used to predict prominence automatically.

6.2 Operational definition of prominence by naive listeners

On the basis of two pilot studies and a larger experiment concerning prominence assignment, an operational definition of prominence was formulated. Prominence is defined as the amount of emphasis attributed by a group of naive native transcribers. For the sake of validity we wanted to keep our listeners as 'naive' as possible as to the task to be accomplished. The marking of prominence in the way we did (binary marking on words by more than one listener) was a useful approach to mark large databases, leaving us with enough detailed information to analyze acoustic and linguistic correlates of prominence. However, the investigation of gradient prominence on an 11-point scale proved too difficult to handle.

In this part of our study, there appeared to be three main discussion items: 1) should we assign prominence at the word or at the syllable level, 2) should we use binary marks or marks on a gradient scale to indicate the degree of prominence and 3) how is the consistency and reliability of the listeners.

6.2.1 Word or syllable prominence marking

An advantage of word prominence assignment is, that it is much easier to perform than syllable-prominence assignment. Additionally, words are more meaningful elements for naive listeners than syllables are. A word is a unit of expression, which has a universal intuitive recognition and meaning by native listeners and speakers. This increases the validity and this makes the labeling intuitively more plausible.
A disadvantage of marking prominence on word level is, however, that one has no detailed information about the identity of the prominent syllable in polysyllabic

words. To circumvent this disadvantages, an alternative approach would be to mark the prominent syllable(s). Using this approach about 250,000 words of the ten-million-words *Corpus Gesproken Nederlands* (short CGN *Spoken Dutch Corpus, http://lands.let.kun.nl/cgn/,* Buhmann et al., 2002) are currently annotated for syllable prominence. This implies that detailed information about the prominence distribution within words, as realized in utterances, will become available. However, the labelers had to be trained to mark syllable prominence according to a detailed protocol. This increases the consistency, but decreases the simplicity and the applicability of the approach.
In our study, it was decided to ask the listener to mark word prominence rather than syllable prominence. This decision was partly based on the results of the pilot experiment as described in section 2.3, which showed that word prominence tends to be assigned more consistently than syllable prominence. For all analyses presented in this study this word-based approach appeared to be detailed enough. Lexical (word) stress was used as a next best approximation to identify the most prominent syllable in those words that were labeled as prominent. For further research it is interesting to investigate whether different strategies were used to mark prominence on the syllable level in comparison to the word level.

6.2.2 Binary or gradient prominence marking

Fant & Kruckenberg (1989), Portele & Heuft (1997), and Grover et al. (1997) used naive listener judgments to define prominence in a detailed way. In their studies, prominence judgments were given for every syllable in the sentence on a very detailed scale. Such a detailed prominence assignment was unpractical in the present study, because it would not allow the annotation of large acoustic databases. Moreover, such detailed information about prominence generally appeared to be unnecessary, even according to the above authors: they all considerably reduced the detailed prominence scales in their further analyses.
In the present study it was decided to ask from each listener a binary prominence mark instead of an n-points gradient prominence scale. Binary marking makes the annotation task easier to perform. This appeared to be a useful approach in our study, whereas the cumulative marks over a group of listeners provided a useful indication for the gradient degree of prominence.
As pointed out before problems remained with application of the cumulative 11-point scale. We have reduced this scale to four prominence classes by means of a hierarchical cluster analysis and even reduced these four classes to a binary prominence distinction. While ten listeners marked the training set, one 'normative' listener was selected to mark the test set. Further research is needed to find out what a useful and necessary detailed range of the prominence scale is. Related to this, questions arose such as what is the relationship of prominence and the linguistic phenomenon of lexical (word) stress in terms of 'stressed' and 'unstressed' syllables, and / or distinction of four degrees 'primary', 'secondary' 'tertiary' and 'weak'. The phonetic notion of pitch accent and the relationship of prominence needs further research to determine if words marked as highly prominent receive

always an accent-lending pitch movement, and whether words being marked as less prominent belong to e.g. 'secondary' stressed words.

6.2.3 Consistency and reliability

We found that listeners are rather consistent and reliable in their prominence judgments. On average the listeners agreed with one another with a Cohen's Kappa of $\kappa = 0.50$. Considering that one had the freedom to mark just one or several words for prominence per sentence, this is a reasonable degree of consistency. Most of the differences between- and within-listeners can be explained by having a different threshold for prominence marking or, in case of the within-listener differences a threshold shift for prominence marking. Training of the labelers and close instructions by the researchers may help to increase the agreement. However, we point out that we aimed at the smallest influence from the researcher as possible, because the interpretation of prominence (e.g. in terms of the number of prominence marks per sentence) has to be defined by the listener. The listener has been asked to 'mark those words that he / she perceived to be pronounced with emphasis', and had to confirm whether he / she understood the task prior to the annotation itself.
Although this has not been investigated, one must assume that the listeners all have used their own strategy for judging prominence on words. It would be interesting for further research to investigate whether the listeners use different strategies to mark prominence.

6.2.4 Concluding remarks

The growing number of publications on prominence shows the great interest in this topic, especially in combination with speech technology. At some point there were even suggestions to change TOBI into a TOBI-lite version in which the number of different pitch accent types would be reduced and a degree of prominence would be added (Wightman & Rose 1999; Wightman et al., 2000). These publications underline that there is a need to come to a good definition of prominence and how to use prominence.

6.3 Lexical / syntactic correlates of prominence

In chapter 3 we described linguistic correlates / determinants of prominence, how they were extracted and how they were used to predict prominence. These predictions were exclusively based on information that is derived automatically from the text. A detailed analysis of the linguistic determinants gives on the one hand insight into the relationships that allow automatic prediction of prominence. On the other hand the analysis gives insight into the linguistic information, namely the expectation of prominence (top-down information) the listener uses.

6.3.1 Individual correlates

After considering many options of the lexical and syntactic correlates of prominence, we drew up the following list of promising candidates:

A) word class;
B) word length;
C) Adjective-Noun combinations;
D) the first content word of the sentence;

A) Our first finding is that word classes can be ordered according to their ability to carry prominence. The following ranking of increasing prominence is found for the sentence material we used: Article, Conjunction, Pronoun, Auxiliary verb, Verb, Numeral, Adverb, Adjective, Noun and Negation. Note that Negations normally belong to the category of function words, which are considered to be less prominent, but which were marked as 'highly prominent' in our speech material. For the other word classes the overall group distinction between function word / content word remains.

B) Secondly, it was found that the word length, expressed in the number of syllables, is a useful correlate of prominence. The metrical weight, referring to the complexity and the length of a word, is related as well. However, we have not investigated whether another component, the 'complexity of a word', is also related to prominence.

C) The third interesting relationship is found between Adjective-Noun combinations and prominence. The Noun in such a combination is generally less prominent than in other combinations.

D) The last striking finding is that the first content word is often a very prominent one in these read-aloud sentences. This may be specific for this type of speech material, but even then such a type of relationship could still be profitable to predict prominence. We did not investigate the possibility of testing the prominence prediction algorithm on utterances with an entirely different grammatical structure (main and sub clauses, questions). It would be interesting for further research to find out how this relationship behaves for other speaking styles and text types. E.g. in a dialogue situation there may be no need to mark the first content word of an utterance, because the contextual situation is clearer than in separate read-aloud sentences.

6.3.2 Prominence prediction on textual input

All these relationships were implemented into an algorithm to predict prominence. Our algorithm initially predicted the degree of prominence on a 4-point scale, which was later reduced to a 2-point scale. The final binary prominence prediction on an

independent test set appeared to be 81.2% correct. The prediction on the test set based on textual input agrees with a Cohen's Kappa of $\kappa = 0.62$ with the marks of the listener who showed the highest consistency on the training set. This agreement was better than the mean agreement between listeners ($\kappa = 0.50$). So, our algorithm for binary prominence prediction from text shows an agreement that is at least similar to that of between-listeners.

6.3.3. Discussion about lexical / syntactic correlates

6.3.3.1 Comparison to the literature

Comparing our results to other research it appeared that we have achieved similar results. A performance of 80-90% correct prediction of pitch accent placement is reported by Hirschberg (1993). She included automatically derived discourse information in her predictions. A result of 82.5% correct prediction of accent placement is reported by Ross & Ostendorf (1996). They used hand-labeled boundary information, but automatically derived Part-of-Speech tags and even topic information to predict accent placement. Vereecken et al. (1998) predict degrees of prominence on a 4-point scale for Dutch with a performance of about 80% correct.
The task to predict different degrees of prominence is difficult and complicated. This is also reflected in the literature mentioned in the introduction and especially in the research of Widera et al. (1997).

6.3.3.2 Method used

The method we used to analyze the linguistic correlates of prominence consisted of the following steps. First, we had a closer look at the linguistic data and tried to find dependencies. Second, we translated these dependencies into simple heuristic rules, and third, these rules were validated on an independent test set. All these steps were performed fully automatically. This makes our findings relevant in two ways a) they show that certain annotation tasks by humans can be simulated by algorithms with similar or better reliability and b) for speech technology applications. The optimization method we used is a heuristic one, although there are more sophisticated techniques available to analyze large databases. Some of these more probabilistic techniques such as classification trees or artificial neural networks may yield higher performance in predicting prominence. However, with such probabilistic techniques it is more difficult to extract specific knowledge about the relationship of perceived word prominence and lexical and syntactic correlates. Such extraction is explicitly possible from the heuristic rules derived in this study.

6.3.3.3 Useful for Text-to-Speech

When the amount of prominence for each word in a sentence could accurately be predicted it would greatly improve the intelligibility, the naturalness and the pleasantness of Text-to-Speech systems. Rules to predict different degrees of prominence, which are solely based on automatically derived textual information, would be very useful for more sophisticated Text-to-Speech systems. Because of the inter-subject differences (see section 2.4.1.2), a perfect synthesis of a prominence contour will be very difficult if ever possible. Furthermore, this is not even necessary, as only one good prominence prediction for all the words in a sentence is needed for speech synthesis purposes.

6.3.3.3.1 No context information available for 'our' read-aloud sentences

In our case, a disadvantage might be that the speech material is not uttered in context. Therefore it was impossible to determine 'focus' and / or 'given' and 'new' information. An advantage could be that the reading of these sentences is a default reading and that the material is also useful for certain technological applications. However, the speech material is designed for speech recognition; for speech synthesis only one professional example speaker is needed. From literature it is known that there are speaker-dependencies (especially gender) of perceived prominence of F_0 peaks (Gussenhoven & Rietveld, 1998). Since the speech material that we used contains a lot of different speakers, these speaker dependencies are averaged out in the analysis results.

6.3.3.3.1 Translation into acoustic properties

A complicating factor for Text-to-Speech systems is the need to translate the different degrees of prominence into proper acoustic values. This is, however, not our main concern in this study. As Hermes (1991) reports, a falling pitch movement is generally perceived as less prominent than a rising pitch movement with the same excursion size measured in ERB's. A first-order normalization for this finding might be possible, but we have to keep in mind that the exact relationship might be complex. Another problem is the detection of when the pitch movement starts as compared to the onset of the vowel. A pitch movement starting late in the vowel has a different effect on the perceived prominence than a pitch movement starting very early (Hermes, 1995).

6.3.3.3.2Testing the prosody

For speech synthesis purposes, actual testing of the prediction of degrees of prominence with the rules developed and described in chapter 3 will be difficult. We made some preliminary attempts in a pilot experiment, but were not very successful. The different degrees of prominence must first be properly translated into acoustic correlates that cannot be limited to certain pitch movements only. This must then be

implemented into an existing synthesizer in order to compare it with a default algorithm. Much research will have to be done concerning this problem.

6.4 Acoustic correlates of prominence

Acoustic features concerning F_0, duration and intensity have been used by us to discriminate between non-prominent and prominent words. A detailed analysis of possible acoustic features, which are extracted from a unit not larger than a word, is performed. The current analysis concentrates on the individual unit (vowel, syllable and word) and did not investigate relative features e.g. looking at the previous or following unit(s). We came to the conclusion that prominence (as assigned by naive listeners) is reflected in the acoustic speech signal of an individual unit. The listener perceived variations in duration, intensity and F_0 and could use it as bottom-up information contributing to the prominence of a word.

6.4.1 Individual acoustic features

Several selected features based on F_0, duration and intensity at syllable, word or sentence level can automatically be extracted from the speech signal. From the twelve acoustic features, the most distinctive single feature for binary prominence prediction appeared to be the range of F_0 measured in semitones. The F_0 range per word showed a better ability to discriminate prominent and non-prominent words than the F_0 range per syllable. The scores for correct prominent / non-prominent classification were 72% and 69% on word and syllable level, respectively. In this study it is found that syllable duration is also a powerful feature for prominence prediction; even a better one than vowel duration. Without any corrections for intrinsic vowel duration and/or the number of phonemes, binary prominence classification using only syllable duration gave about 71% correct. On the basis of vowel duration as a single feature, a correct binary prominence prediction of 66% was reached. Vowel intensity used as the only input feature to classify prominent and non-prominent words gave results of about 67% correct. This result indicates that vowel intensity is also an important cue for prominence.

6.4.2 Prominence prediction on acoustic input

As observed in the previous section, several individual features showed good results. However, a neural-net classification with all twelve selected features performed better. On the independent test set an appropriately trained neural net performs at 79% correct classification. Comparing the agreement of, on the one hand, the predicted prominence marks between those of the listener in the test set (Cohen's Kappa $\kappa = 0.57$, see table 5.3) with, on the other hand, the average agreement between listeners on the training data (Cohen's Kappa $\kappa = 0.50$), it can be concluded that the neural net prediction is at least as good as the naive listeners' performance.

6.4.2.1 Complexity

From the analyses presented in chapter 5 it also became clear that the relationship between the acoustic features and prominence is not simply linear, but sometimes rather complex. This conclusion justifies and probably partly explains the fact that the use of a hidden layer substantially improves the recognition performance of the neural networks, in comparison with an LDA. What, however, the structure is of this complexity needs further investigations because of the multitude of confounding factors. E.g. how long a given syllable had to be in combination with which changes in pitch is an interesting research question.

6.4.3. Discussion about acoustic correlates

Various points of our approach that may require further discussion are: firstly a comparison to the literature, secondly the method used, thirdly the HMM-alignment, fourthly the applied normalization and lastly, the strictly separate use of linguistic and acoustic features.

6.4.3.1 Comparison to the literature

Just as for the prediction of prominence with textual input, our prediction method with acoustic input seems to produce results comparable to those reported in the literature. The statistical and / or brute force method used by the authors mentioned below aimed at high recognition rates whereas our approach aimed at different perspectives of acoustics and classification providing insight in phonetic questions as well. Our prominence classification results were achieved using solely twelve selected acoustic input features. The comparison is not completely fair as it concerns different methods for training and testing and differences in speech material used. Kießling (1996) reported a recognition rate of 83% correct accent classification for the VERBMOBIL-speech data while using also textual features such as identity of the vowel. Kompe et al. (1995) report classification rates of 95.6% correct for the ELRA corpus. This corpus contains read-aloud sentences with a simple grammatical structure. Wightman & Ostendorf (1994) reached 83% correct using hand-labeled boundary features and Silipo & Greenberg (1999) classified stressed and unstressed syllables with 80% and 77% correct, respectively.

6.4.3.2 Method used

Just as in section 6.3, where we discussed the lexical / syntactic approach, we will now discuss the acoustic approach. Following a detailed analysis of some of the most promising acoustic correlates (correlates that were promising on an individual basis), we selected those features that seemed to have the greatest potential to predict prominence using a multidimensional classification. This pre-selection reduced the number of features for the classification task. In this study it was decided to use simple feed-forward neural networks for predicting prominence. As

said before, sophisticated techniques such as CART, and more complex self-learning neural networks are available, but we decided to put emphasis on a full understanding of the classification process. Within the classification process both the individual features as well as various combinations were tested for their contribution to discriminate prominence.

6.4.3.3 HMM-alignment

In our approach a human-made orthographic transcription was available to do the HMM-alignment. This alignment is used to automatically obtain the segmentation (in terms of phoneme and word boundaries) of the utterances. In realistic speech recognition situations such a precise and correct word-level transcription of what has been said is generally not available. The only available transcription in such a case is the result of automatic speech recognition, which introduces additional errors into the automatic alignment. Furthermore, the automatic HMM-alignment we used introduces, by its very nature, errors and problems to the resulting segmentation. The automatic segmentation used in this study was based on Dutch standard pronunciation, although the pronunciation of some speakers showed regional variants. We did not investigate the precise consequences of this in any detail, as in speech technology in general one has to cope with similar conditions.

6.4.3.4 Normalizations

Apart from possible segmentation errors ascribed to the assumption of a standard pronunciation, other segmentation errors are unavoidable in this automatic procedure. This may cause fewer problems for measurements in larger segments, such as syllables and words, than for measurements in smaller segments, such as vowels and consonants. Normalizations concerning intrinsic vowel duration and intrinsic vowel intensity were conducted at such small units, namely vowels. Maybe this fact and the large speaker variability explain why these normalizations on the level of small units did not substantially improve the ability to discriminate between prominent and non-prominent words. Analyses on sentence speaking rate were also conducted in this study. Speaking rate is reported to influence the duration of vowels and syllables. Although an effect of sentence speaking rate on vowel and syllable duration is indeed shown in chapter 4 (see section 4.2.2.3 and figure 4.14), the normalization for speaking rate that we applied, did not improve prominence classification (section 5.2.7). Further research is required to investigate the precies effect of normalizations. We tried overall-normalizations (intrinsic vowels duration, intrinsic vowel intensity, sentence speaking rate) with no effect on the prominence classification when used separately. However, the approach and the speech material may have triggered this negative result. Other research shows that putting the vowel / syllable / word in larger context (e.g. taking the previous and next syllable and /or word into account) increases the classification results. Taking into account these relative features in a detailed analysis may help to get more insight into the relative character of prominence.

6.4.3.5 Separate use of linguistic and acoustic features

The strictly separate use of linguistic and acoustic features may be disadvantageous for syllable duration. Syllable duration might also be influenced by linguistic information, such as word class. Content words often consist of more complex syllables than function words; this means that syllables of content words contain more phonemes. This higher complexity may result in longer syllable duration. So, there might be a relationship between syllable duration and the content word / function word distinction. Fant & Kruckenberg (1999) noticed that syllable duration was the most robust correlate of prominence. Furthermore, we have of course to keep in mind that pitch extraction is not always error free (octave errors). An example of such an error was demonstrated in chapter 4, figure 4.4.

6.5 Future research

In this section, first some suggestions for the use of other promising features of prominence are given. This concerns for instance the relationship between word prominence and spectral quality. Next, a few speech-technological applications are discussed and finally, a number of remarks are made about combining linguistic and acoustic features.

6.5.1 Promising features of prominence

A possibly promising correlate, such as the distance between two prominent words in a sentence (speech rhythm, stress clash), was not investigated. We do not exclude the possibility, however, that another prominent word may be required to follow the first prominent word after a period of time in a sentence. Information concerning the text structure, for example, whether or not words convey given / new information or boundary information, as used in Wightman & Ostendorf (1994), could not be assigned and used in our analyses. Pragmatic and semantic information could not be derived automatically and was thus not available, and is most probably less relevant for isolated sentences. For future research on the improvement of speech synthesis based on prominence, the use of such more elaborate information about text structure could be helpful.
Spectral quality was also excluded from our analyses. We took this decision since consistent spectral properties are difficult to measure automatically. Further research will be required to measure spectral quality reliably and to learn more about its relationship with prominence. One of the issues to be solved is a phoneme-based normalization of the spectral quality measure.

6.5.2 Speech-technological applications

In this study several suggestions have been made about how prominence could be used in speech-technological applications. The suggestions include: the improvement of the intonation of speech synthesis, a word-by-word prominence

indicator, and sentence disambiguation. However, such implementations require further research.

Predicting prominence from textual input could help to improve speech synthesis. In the presently popular and very promising approaches of (variable) unit-based concatenative speech synthesis (Klabbers, 2000; Stöber et al., 1999; Wightman et al. 2000), the prediction of prominence (from text) is of crucial importance for the pleasantness and naturalness of the synthesized speech signal. The text-derived prominence labels will have to be matched with the prominence labels in the annotated speech database that contains the segments to be concatenated. A search algorithm is supposed to select the optimally matching segments and to concatenate them. So, a translation from prosodic labels to acoustic parameters is not always needed: the (larger) speech units in the database already contain most of that information, although sometimes additional signal adaptations are required. However, for diphone synthesis a translation from prominence labels to acoustical parameters will almost always be necessary. How exactly prominence labels that are predicted from text can be translated into acoustic features is beyond the scope of this study and needs further research. The exact relationship between prominence and information retrieval (focus, contrast, topic, etc.) has not been investigated either.

As mentioned earlier in this study, prosody is so far hardly used in present day speech recognition. We suggested two applications: a word-by-word prominence indicator and an instrument to disambiguate the meaning of an ambiguous sentence. Although, sentence disambiguation has already been a topic for research (Batliner et al., 1998, for instance within the German Verbmobil project), a running prominence indicator for speech recognition is a new idea that still awaits application. Ida & Yamasaki (1998) show improvements for keyword spotting as used in speech recognition based on prosodic information. Knowledge about or estimation of the prominence of a word during the recognition process can provide islands of reliability or can point out the importance of a word. Wang & Seneff (2001) and van Kuijk & Boves (1999) used lexical stress determined through lexical look-up to improve speech recognition, but the improvement they could achieve was negligible. Another example of using prosody is given in Taylor et al. (1998). They described how prosody helped to constrain speech recognition in a dialogue environment. Positive results in this area have been reported in recent papers by Hirschberg and Swerts (1998), and by Wang (2001).

6.5.3 Combination of linguistic and acoustic information

Combining acoustic and linguistic features may improve prominence prediction as shown by Vereecken et al. (1998). In speech synthesis only textual information is available, whereas in speech recognition only acoustical information is available. However, for the annotation of a speech corpus, usually both acoustical and textual information are available. For the large 10 million words CGN Corpus it is the intention to annotate prominence of 250,000 words by hand. Automatic labeling

procedures can substantially help to consistently annotate large databases. These procedures can also be useful to improve the quality of concatenative speech synthesis.
Linking the linguistic and the acoustic features can lead to more reliable recognition / prominence classification rates, and furthermore it will also be very useful for automatic annotation. Further tests and further research will be required to investigate whether automatic annotation will be possible, especially if it concerns a lot of different speaking styles and different speakers. However, the detailed analyses of prominence as presented in this study, as well as the acoustic and linguistic correlates, hopefully do provide much information and various suggestions for further improvements of speech technology.

Form a more scientific viewpoint it is interesting to know more about the recognition process of prominence in general. That prominence is reflected in the speech signal of individual units was shown in the present study. How far the listener uses this bottom-up information to match his expectations of prominence on the basis of linguistic knowledge of his languages (top-down) is an interesting research topic.
That prominence is reflected in the lexical and syntactic knowledge was also shown in the present study. However, to which end the listener uses 'our' correlates and how the resulting expectation is matched with the bottom-up information related to the speech signal is beyond the scope of this study, however, very interesting for future research.

This study underlines that prominence is reflected in the acoustic and the linguistic domain, and that a binary prominence prediction with a selected set of relatively simple features can lead to a performance similar to that of naive listeners.

Appendix 2.1: Correspondence matrix assigning prominence without pitch movements.

Table A 2.1: Correspondence matrix for the prominence word scores of the perception experiment without pitch movements, and the results for the same 30 sentences from the regular word perception experiment.

		Word experiment without pitch movements									
	Listener	0	1	2	3	4	5	6	7	8	Total
Word experiment	0		25	10	0	0	0	0	0	0	153
	1	6		5	5	2	0	0	0	0	26
	2	2	4		2	1	0	0	0	0	11
	3	2	7	2		2	2	0	0	0	20
	4	2	3	4	4		0	1	0	0	16
	5	1	1	9	3	7		1	0	0	25
	6	0	1	4	3	4	4		0	0	16
	7	0	2	2	3	6	3	4		0	20
	8	0	0	1	1	3	7	7	4		25
	Total	131	51	39	26	27	19	13	4	2	312

Appendix 2.2: NIST header.

The header of a NIST file (the file format used for storing the speech material) contains among other information, information about the speaker, about the recording and about the assessment.

```
NIST_1A
  1024
speaking_mode -s4 read
caller_id -s8 tf1002zh
age -s2 48
gender -s6 female
region -s12 Zuid-Holland
education_level -s1 2
cordless_phone -s2 no
sheet_identifier -s4 7419
recording_date -s6 931228
recording_time -s6 150106
database_id -s12 POLYPHONE-NL
database_version -s3 1.0
microphone -s9 telephone
sample_rate -i 8000
sample_count -i 50816
channel_count -i 1
sample_n_bytes -i 1
sample_sig_bits -i 8
response_category -s26 phonetically_rich_sentence
prompt_text -s16 14. Lees de zin:
sheet_text -s53 We zijn constant bezig de drugsrunners te bestrijden.
transliteration -s52 we zijn constant bezig de drugsrunners te bestrijden
assessment -s2 OK
sample_coding -s27 alaw,embedded-shorten-v1.09
sample_byte_format -s1 1
sample_checksum -i 10929
end_head
```

Appendix 2.3: Demographic data of the training set.

Table A 2.3: The total numbers and percentages of the geographical origin, sex, age and education level of the speaker from the training set data.

Region	Number	Percentage
Zeeland	92	7.4
Noord-Holland	202	16.2
Zuid-Holland	203	16.3
Noord-Brabant	177	14.2
Gelderland	154	12.4
Utrecht	102	8.2
Limburg	68	5.5
Overijssel	72	5.8
Friesland	30	2.4
Drenthe	87	7.0
Groningen	45	3.6
Flevoland	12	1.0
Total	1244	100

		Number	Percentage
Sex	Male	631	50.7
	Female	613	49.3
Age	< 20	35	2.8
	20 - 30	287	23.1
	30 - 40	375	30.1
	40 - 50	295	23.7
	50 - 60	144	11.6
	60 - 70	83	6.7
	> 70	25	2.0
Education level	1	52	4.2
	2	652	52.4
	3	540	43.4

Appendix 2.4: Demographic data of the test set.

Table A 2.4: Table A 2.3 but now for the test set.

Region	Number	Percentage
Zeeland	27	2.7
Noord-Holland	231	23.1
Zuid-Holland	203	20.3
Noord-Brabant	100	10.0
Gelderland	102	10.2
Utrecht	82	8.2
Limburg	29	2.9
Overijssel	35	3.5
Friesland	53	5.3
Drenthe	42	4.2
Groningen	71	7.1
Flevoland	25	2.5
Sum	1000	100

		Number	Percentage
Sex	Male	419	41.9
	Female	581	58.1
Age	< 20	21	2.1
	20 - 30	243	24.3
	30 - 40	293	29.3
	40 - 50	218	21.8
	50 - 60	125	12.5
	60 - 70	56	5.6
	> 70	44	4.4
Education level	1	41	4.1
	2	505	50.5
	3	454	45.4

Appendix 2.5: Instructions for the listening experiment.

Beste proefpersoon,

Je krijgt in het volgende luisterexperiment in totaal 550 verschillende zinnen te horen. Deze 550 zinnen zijn verdeeld over 4 sessies. Sessie 1 en 2 bestaan uit 150 en sessie 3 en 4 uit 125 zinnen.

Je moet aangeven welk woord of woorden in de zin met NADRUK zijn uitgesproken.

Dit kun je doen door met de linker muistoets op het knopje onder het woord te drukken, je kunt corrigeren door nog een keer op het knopje te drukken. Heb je dit gedaan druk dan met de muis op het knopje met de tekst 'Klaar', dan begint de volgende zin. Je kunt iedere zin maximaal 3 keer beluisteren, als je al eerder klaar bent kun je op 'Klaar' drukken. Een sessie duurt ca. 45 minuten.

Wilt je eerst hieronder zowel je naam invullen als aanklikken en aangeven om welke sessie het gaat.
Als je dit gedaan hebt kun je op start drukken.

Alvast bedankt voor je medewerking.

Appendix 3.1: Overview of predicted prominence marks and lexical and syntactic correlates.

Table A 3.1: Overview of the groups of words receiving from 0 to 4 prominence marks for the six different sets of rules. For more details see chapter 3, section 3.3.6.

	0 marks	1 mark	2 marks	3 marks	4 marks
Set A	-mono FW	-poly FW -mono V, Adv -mono combi N	-poly V, Adv -mono N, Adj, Num, Neg -poly combi N	-poly N, Adj, Num, Neg -first V, Adv	-first poly N, Adj, Num, Neg
Set B	-FW	-poly Pron -mono V, Adv	-poly V, Adv -combi N -first mono V, Adv	-N, Adj, Num, Neg -first poly v, Adv	-first N, Adj, Num, Neg
Set C	-Art, Aux -mono Conj, Prep, Pron	-poly Conj, Prep, Pron -mono V, Adv -mono combi N	-Poly V, Adv -mono N -poly combi N -first mono V, Advs	-poly N -Adj, Num, Neg -first V, Adv	-first poly N -first Adj, Num,
Set D	-mono FW	-poly FW -poly V, Adv -combi mono N	-poly V, Adv -mono N, Adj, Num, Neg -combi mono N	-poly N, Adj, Num -first poly V, Adv first mono N, Adj, Num, Neg	-first poly N, Adj, Num
Set E	-FW (not poly Pron)	-poly Pron -mono V, Adv	-poly V, Adv -combi N	-Noun, Num, Adj, Neg -first poly N, Adv	-first N, Num, Adj, Neg
Set F	-Art, Aux -mono Conj, Prep, Pron	-poly Conj, Prep, Pron -mono V, Adv -combi mono N	-poly V, Adv -mono N -combi poly N	-poly N -Adj, Num, Neg -first poly V, Adv -first mono N	-first poly N -first Adj, Num, Neg

Appendix 4.1: SAMPA symbols.

Symbols for the phonemes and for non-speech sounds (SAMPA-like):

Non speech			
n=	big noise		
m=	mouth noise		
p=	pause		
a=	breath		
sil	silence		

Vowels (all vowels, except schwa, may occur in stressed position, indicated by *)

i	long	*niet*	*not*
u	long	*hoe*	*what (how)*
y	long	*u*	*you*
e:	long	*geen*	*not, no*
a:	long	*naam*	*name*
o:	long	*wonen*	*to live*
Q:	long	*kleur*	*color*
E	short	*werk*	*work*
A	short	*dag*	*hello, day*
O	short	*op*	*at*
Y	short	*nummer*	*number*
I	short	*in*	*in*

Diphthongs

Au	*oud*	*old*
Ei	*hij*	*he*
9y	*uit*	*from*

Schwa

@	*geven*	*to give*

Consonants

L	*bal*	*bal*
l	*land*	*country*
N	*lang*	*long*
m	*maand*	*month*
n	*nee*	*no*
R	*reis*	*journey*
r	*reis*	*journey*
j	*jaar*	*year*
f	*fiets*	*bike*
v	*veel*	*much, many*
w	*werk*	*work*
s	*snel*	*fast*
z	*zoon*	*son*
h	*hebben*	*to have*
S	*sjaal*	*scarf*
Z	*bagage,*	*baggage*
x	*groen*	*green*
p	*punt*	*point*
b	*boek*	*book*
t	*taal*	*language*
d	*dochter*	*daughter*
k	*kinderen*	*children*
G	*goal*	*goal*

Appendix 5.1: Scaling values.

Table A 5.1: This table presents the extremes (99% percentile and the 1% percentile) on which the acoustic features used as input are scaled, so that each value lies between one and zero. Outliers (values beyond these two borders) are treated as being 0 or 1, respectively.

Feature	99% percentile	1% percentile
Vowel duration	0.23	0.03
Vowel duration normalized	3.30	-1.72
Speaking rate	0.86	-0.55
Vowel intensity corrected	84.70	60.60
Vowel intensity normalized	1.83	-2.97
Sentence intensity	73.60	65.50
Syllable duration	0.47	0.06
Median F_0	27.70	-9.00
Range F_0 syllable	12.00	0.10
Median F_0 corrected	15.60	-15.00
Median F_0 sentence	20.02	-2.00
Range F_0 word	13.00	0.13

BIBLIOGRAPHY

Altenberg, B. (1987). 'Prosodic patterns in spoken English: Studies in the correlation between prosody and grammar for text-to-speech conversation', *Lund Studies in English,* 76, Lund University Press, Lund.

Baart, J. (1987). *Focus, Syntax and Accent Placement. Towards a rule system for the derivation of pitch accent patterns in Dutch as spoken by humans and machines,* Doctoral dissertation, University of Leiden.

Bagshaw, P.C. (1993). 'An investigation of acoustic events related to sentential stress and accents, in English', *Speech Communication,* 13, 333-342.

Batliner, A., Kompe, R., Kießling, A., Nöth, E., Niemann, H. & Kilian, U. (1995). 'The prosodic marking of phrase boundaries: Expectations and results', In: A.J. Rubio Ayuso & J.M. López Soler, (Eds.), *Speech Recognition and Coding. New Advances and Trends*, Springer-Verlag, Berlin, 89-92.

Batliner, A., Warnke, V., Nöth, E., Buckow, J., Huber, R. & Nutt, M. (1998). 'How to label accent position in spontaneous speech automatically with the help of syntactic-prosodic boundary labels', In: *bmb+f Bundesministerium Bildung, Wissenschaft, Forschung und Technologie, Verbmobil Report 228,* 1-42.

Batliner, A., Buchow, J., Huber, R., Warnke, V., Nöth, E. & Niemann, H. (1999). 'Prosodic feature evaluation: brute force or well designed?', *Proc. ICPhS'99*, San Francisco, 3, 2315-2318.

Batliner, A., Nutt, M., Warnke, V., Nöth, E., Buckow, J., Huber, R. & Niemann, H. (1999). 'Automatic annotation and classification of phrase accents in spontaneous speech', *Proc. Eurospeech'99*, Budapest, 1, 519-522.

Bergem, D.R. van (1993). 'Acoustic vowel reduction as a function of sentence accent, word stress, and word class', *Speech Communication,* 12, 1-23.

Boersma, P. & Weenink, D. (1996). PRAAT: A system for doing phonetics by computer, *Report of the Institute of Phonetic Sciences of the University of Amsterdam* 132, (www.praat.org).

Bolinger, D.L. (1958). 'A theory of pitch accent in English', *Word,* 14, 109-149.

Bolinger, D.L. (1972). 'Accent is predictable (if you're a mind-reader)', *Language,* 48, 633-644.

Bolinger, D.L. (1989). *Intonation and its use: melody in grammar and discourse*, Stanford University Press, London, 1989.

Bosch, L.F.M. ten (1993). 'On the automatic classification of pitch movements', *Proc. Eurospeech'93*, Berlin, 2, 781-784.

Buhmann, J., Vereecken, H., Fackrell, J., Martens, J-P. & Coile, B. van (2000). 'Data driven intonation modelling of 6 languages', *Proc. ICSLP'00*, Peking, 3, 179-182.

Buhmann, J., Caspers, J., Heuven, V.J.J.P. van, Hoekstra, H., Martens J-P. & Swerts, M. (2002). 'Annotation of prominent words, prosodic boundaries and segmental lengthening by non-expert transcribers in Spoken Dutch Corpus', *Proc. LREC'02,* Las Palmas, 779-785.

Bulyko, I. & Ostendorf, M. (1999). 'Predicting gradient F_0 variation: pitch range and accent prominence', *Proc. Eurospeech'99*, Budapest, 4, 1819-1822.

Cambier-Langeveld, T. (2000). *Temporal marking of accent and boundaries*, Doctoral dissertation, University of Amsterdam.

Campbell, N. (1995). 'Prosodic influence on segmental quality', *Proc. Eurospeech'95*, Madrid, 2, 1011-1014.

Chomsky, N. & Halle, M. (1968). *The sound pattern of English*, Harper & Row, New York.

Cohen, J. (1960). 'A coefficent of agreement for nominal scales', *Educational and Physological Measurements,* 20, 37-46.

Cucchiarini, C., Binnenpoorte, D. & Goddijn, S. (2001). 'Phonetic transcriptions in the spoken Dutch corpus: how to combine efficiency and good transcription quality', *Proc. Eurospeech'01*, Aalborg, 1, 1679-1682.

Cutler, A., Dahan, D. & Donselaar, W. van (1997). 'Prosody in the comprehension of spoken language: a Literature Review', *Language and Speech*, 40 (2), 141-201.

Cutler, A. (1984). 'Stress and accent in language production and understanding', In: D. Gibbon, H. & Richter (Eds.), *Intonation, accent and rhythm, studies in discourse phonology*, de Gruyter, Berlin, 77-90.

Damhuis, M., Boogaart, T., 't Veld, C. in, Versteijlen, M., Schelvis, W., Bos, L. & Boves, L. (1994). 'Creation and analysis of the Dutch Polyphone Corpus', *Proc. ICSLP'94,* Yokohama, 4, 1803-1806.

Daelemans, W., Zavrel, J., Berck, P. & Gillis, S. (1996). 'MBT: A memory based Part-of-Speech Tagger-Generation', In: E. Ejerhed & I. Dagan (Eds.), *Proc. Fourth Workshop on very Large Corpora*, 14-27.

Dirksen, A. & Quené, H. (1993). 'Prosodic analysis: The next generation', In: V.J.J.P. van Heuven & L.C.W. Pols (Eds.), *Analysis and synthesis of speech: Strategic research towards high-quality TTS generation*, Mouton de Gruyter, 131-144.

Donzel, M.E. van (1999). *Prosodic aspects of information structure in discourse,* Doctoral dissertation, University of Amsterdam.

Donzel, M.E. van & Koopmans-van Beinum, F.J. (1995). 'Prominence judgements and textual structure in discourse', *Proc. Institute of Phonetic Sciences of the University of Amsterdam,* 19, 11-23.

Eisen, B. & Tillmann, H.G. (1992). 'Consistency of judgements in manual labelling of phonetic segments: The distinction between clear and unclear cases', *Proc. ICSLP'92*, Banff, 2, 871-874.

Fackrell, J., Vereecken, H., Martens, J-P. & Coile, B. van (1999). 'Multilingual prosody modelling using cascades of regression trees and neural networks' *Proc. Eurospeech'99*, Budapest, 4, 1835-1838.

Fackrell, J., Vereecken, H., Buhmann, J., Martens, J-P. & Coile, B. van (2000). 'Prosodic variation with test type', *Proc. ICSLP 2000*, Bejing, 3, 231-234.

Fant, G. & Kruckenberg, A. (1989). 'Preliminaries to the study of Swedish prose reading and reading style', *Speech Trans. Q. Prog. Stat. Rep.*, 2/1989 KTH, Stockholm, 1-83.

Fant, G. & Kruckenberg, A. (1999). 'Prominence correlates in Swedish prosody', *Proc. ICPhS'99*, San Francisco, 3, 1749-1752.

Fosler-Lussier, E., Greenberg, S. & Morgan, N. (1999). 'Incorporating contextual phonetics into automatic speech recognition', *Proc. ICPhS'99*, San Fransisco, 1, 611-614.

Fosler-Lussier, E. & Morgan, N. (1999). 'Effects of speaking rate and frequency on pronunciations in convertional speech' (sic), *Speech Communication*, 29, 137-158.

Frid, J. (2001). 'Prediction of intonation patterns of accented words in a corpus of read Swedish news through pitch contour stylization', *Proc. Eurospeech'01*, Aalborg, 2, 915-918.

Fry, D.B. (1958). 'Experiments in the perception of stress' *Language and Speech*, 1, 126-152.

Grabe, E., Nolan, F. & Farrar, K.J. (1998). 'IViE - A comparative transcription system for intonational variation in English', *Proc. ICSLP'98*, Sydney, 4, 1259-1262.

Granström, B., House, D. & Lundeberg, M. (1999). 'Prosodic cues in multimodal speech perception', *Proc. ICPhS'99*, San Francisco, 1, 655-658.

Grover, C. & Terken, J. (1995). 'The role of stress and accent in the perception of rhythm', *Proc. ICPhS'95*, Stockholm, 4, 356-359.

Grover, C., Heuft, B. & Coile, B. van (1997). 'The reliability of labeling word prominence and prosodic boundary strength', *Proc. ESCA Workshop on Intonation*, Athens, 165-168.

Grover, C., Fackrell, J., Vereecken, H., Martens, J-P. & Coile, B. van (1998). 'Designing prosodic databases for automatic modelling in 6 languages', *Proc. third ESCA/COCOSDA Workshop on Speech Synthesis*, Jenolan Caves, 93-98.

Gussenhoven, C.H. (1984). *On the grammar and semantics of sentence accents*, Doctoral dissertation, University of Nijmegen, Foris Publication, Dordrecht.

Gussenhoven, C.H. (1985). 'On the relation between pitch excursion size and prominence', *Journal of Phonetics*, 13, 299-308.

Gussenhoven, C.H., Rietveld, A., Rump, H.H. & Terken, J. (1997). 'The perceptual prominence of fundamental frequency peaks', *J. Acoust. Soc. Am.*, 102 (5), 3009-3022.

Gussenhoven, C.H. & Rietveld, T. (1998). 'On the speaker-dependence of the perceived prominence of F_0 peaks', *Journal of Phonetics* 26, 371-380.

Halliday, M.A.K. (1967). 'Notes on transitivity a theme in English', part II., *Journal of Linguistics*, 3, 199-244.

't Hart, J., Collier, R. & Cohen, A. (1990). *A perceptual study of intonation*, University Press, Cambridge.

Heldner, M., Strangert, E. & Deschamps, T. (1999). 'A focus detector using overall intensity and high frequency emphasis', *Proc. ICPhS'99*, San Francisco, 3, 1491-1494.

Helsloot, C.J. (1995). *Metrical Prosody. A template-and-constraint approach to phonological phrasing in Italian,* Doctoral dissertation, University of Amsterdam.

Helsloot, C.J. & Streefkerk, B.M. (1998), 'Perceived prominence and the metrical-prosodic structure of Dutch sentences', *Proc. Institute of Phonetic Sciences of the University of Amsterdam,* 22, 97-113.

Hermes, D.J. (1991). 'Prominence caused by rising and falling pitch movements with different position in the syllable', *IPO Annual Progress Report,* 26, 17-28.

Hermes, D.J. & Rump, H.H. (1994). 'Perception of prominence in speech intonation induced by rising and falling pitch movements', *J. Acoust. Soc. Am.,* 96 (1), 83-92.

Hermes, D.J. (1995). 'Timing of pitch movements and accentuation of syllables', *IPO Annual Progress Report,* 30, 38-44.

Hess, W., Batliner, A., Kießling, A., Kompe, R., Nöth, E., Petzold, A., Reyelt, M. & Strom, V. (1996). 'Prosodic modules for speech recognition and understanding in VERBMOBIL', In: Y. Sagisaka, N. Campbell & N. Higuchi (Eds.), *Computing Prosody. Approaches to a Computational Analysis and Modelling of the Prosody of Spontaneous Speech,* Springer-Verlag, New York, 361-379.

Heuft, B., Streefkerk, B.M. & Portele, T. (1996). 'Evaluierung der automatischen Parametrisierung von Grundfrequenzkonturen', *Proc. Elektronische Sprachsignalverarbeitung,* 7, Berlin, 170-175.

Heuven, V.J.J.P. van (1994). 'What is the smallest prosodic domain?' In: P.Keating (Eds.) *Papers in Laboratory Phonology III: phonological structure and phonetic form,* Cambridge University Press, London, 76-98.

Hieronymus, J.L. (1989). 'Automatic sentential vowel stress labelling', *Proc. Eurospeech'89,* Paris, 1, 226-229.

Hirschberg, J. (1990). 'Accent and discourse context: Assigning pitch accent in synthetic speech', *Proc. 8th National Conference on AI,* Menlo Park, 952-957.

Hirschberg, J. (1993). 'Pitch accent in context: predicting intonational prominence from text', *Artificial Intelligence,* 63, 305-340.

Hirschberg, J. & Swerts, M. (1999). 'Prosody and conversation: an introduction', *Language and Speech,* 41:3/4, 229-233.

Hirschberg, J. & Rambow, O. (2001). 'Learning prosodic features using a tree representation', *Proc. Eurospeech'01,* Aalborg, 2, 1175-1178.

Ida, M. & Yamasaki, R. (1998). 'An evaluation of keyword spotting performance utilizing false alarm rejection based on prosodic information', *Proc. ICSLP'98,* Sydney, 3, 803-806.

Johnson S. (1967). 'Hierarchical Cluster Schemes', *Psychometika,* 38, 241-254.

Kießling, A., Kompe, R., Batliner, A., Niemann, H. & Nöth, E. (1994). 'Automatic labeling of phrase accents in German', *Proc. ICSLP'94,*Yokohama, 1, 115-118.

Kießling, A., Kompe, R., Niemann, H. & Nöth, E. (1994). 'Detection of phrase boundaries and accents', In: H. Niemann, R. de Mori & G. Hanrieder (Eds.), *Progress and Prospects of Speech Research and Technology,* Infix, 266-269.

Kießling, A. (1996). *'Extraktion und Klassifikation prosodischer Merkmale in automatischer Sprachverarbeitung*, Berichte aus der Informatik, Aachen.

Kießling, A., Kompe, R., Batliner, A., Niemann, H. & Nöth, E. (1996). 'Classification of boundaries and accents in spontaneous speech', *Proc. CRIM/FORWISS Workshop*, Montreal, 104-113.

Klabbers, E. (2000). *Segmental and prosodic improvements to speech generation*, Doctoral dissertation, University of Eindhoven.

Kompe, R., Kießling, A., Niemann, H., Nöth, E., Schukat-Talamazzini, E., Zottmann, A. & Batliner A. (1995). 'Prosodic scoring of word hypothesis graphs', *Proc. Eurospeech'95,* Madrid, 2, 1333-1336.

Koopmans-van Beinum, F.J. (1980). *Vowel contrast reduction, an acoustic and perceptual study of Dutch vowels in various speech conditions*, Doctoral dissertation, University of Amsterdam, Academische Pers B.V., Amsterdam.

Kraayeveld, J., Rietveld, A.C.M. & Heuven, V.J.J.P. van (1991). 'Speaker characterization in Dutch using prosodic parameters', *Proc. Eurospeech'91*, Genova, 2, 427-430.

Kraayeveld, J., Rietveld, A.C.M. & Heuven, V.J.J.P. van (1993). 'Speaker specificity in prosodic parameters', *ESCA'93 Workshop on Prosody, Working papers 41, Dept of Linguistics and Phonetics,* Lund, 264-267.

Kuijk, D. van & Boves, L. (1999). 'Acoustic characteristics of lexical stress in continuous telephone speech', *Speech Communication,* 27, 95-111.

Ladd, D.R., Verhoeven, J. & Jacobs, K. (1994). 'Influence of adjacent pitch accents on each other's perceived prominence: two contradictory effects', *Journal of Phonetics* 22, 87-99.

Ladd, D.R. (1996). *Intonational phonology*, Cambridge, Univerisity Press.

Ladd, D.R. (1980). *The structure of intonational meaning: Evidence from English*, Bloomington, Indiana University Press.

Lea, W.A. (1980). 'Prosodic aids to speech recognition', In: W.A. Lea (Eds.), *Trends in Speech Recognition*, Prentice Hall, INC., Englewood Cliffs, New Jersey.

Lehiste, I. (1970). *Suprasegmentals*, Cambridge, Mass. MIT Press.

Lehiste, I. & Peterson, G.E. (1959). 'Vowel amplitude and phonemic stress in American English', *J. Acoust. Soc. Am.,* 31, 428-435.

Lippmann R.P. (1987). 'An introduction to computing with neural nets', *IEEE ASSP Magazine,* 4-22.

Maghbouleh, A. (1998). 'ToBI accent type recogniton', *Proc. ICSLP'98*, Sydney, 3, 639-642.

Mozziconacci, S. (1998). *Speech variability and emotion, production and perception*, Doctoral dissertation, University of Eindhoven.

Nöth, E., Niemann, H. & Schmölz, S. (1988). 'Prosodic features in German speech: Stress assignment by man and machine', In: H. Niemann, M. Lang, & G. Sagerer, *Recent Advances in Speech Understanding and Dialog Systems,* NATO ASI Series F, Springer-Verlag, Berlin, 46, 101-106.

Nooteboom, S.G. & Kruyt, J.G. (1987). 'Accents, focus distribution, and the perceived distribution of given and new information', *J. Acoust. Soc. Am.*, 82, 1512-1524.

Petzold, A. (1995). 'Strategies for focal accent detection in spontaneous speech', *Proc. ICPhS'95,* Stockholm, 3, 672-675.

Pfitzinger, H.R. (1999). 'Local speech rate perception in German speech', *Proc. ICPhS'99*, San Francisco, 2, 893-896.

Pierrehumbert, J. (1980). *The phonology and phonetics of English intonation*, MIT Linguistics, Doctoral dissertation, Bloomington, Indiana.

Pijper, J.R. de & Sanderman, A.A. (1994). 'On the perceptual strength of prosodic boundaries and its relation to suprasegmental cues', *J. Acoust. Soc. Am.*, 96, 2037-2047.

Portele, T. & Heuft, B. (1995). 'Two kinds of stress perception', *Proc. ICPhS'95*, Stockholm, 1, 126-129.

Portele, T. & Heuft, B. (1997). 'Towards a prominence-based speech synthesis system', *Speech Communication,* 21, 61-72.

Portele, T. (1998). 'Perceived prominence and acoustic parameters in American English', *Proc. ICSLP'98*, Sydney, 3, 667-670.

Portele, T. (1999). 'A perceptual motivated intonation model for German', *Proc. ICPhS'99*, San Francisco, 2, 949-952.

Press, W.H., Teukolsky, S.A., Vetterling, W.T. & Flannery, B.P. (1992). '*Numerical recipes in C: the art of scientific computing*', Second Edition, Cambridge, University Press.

Price, P.J., Ostendorf, M., Shattuck-Hufnagel, S. & Fong, C. (1991). 'The use of prosody in syntactic disambiguation', *J. Acoust. Soc. Am.*, 90 (6), 2956-2969.

Quené, H. & Kager, R. (1993). 'Prosodic sentence analysis without exhaustive parsing', In: V.J.J.P. van Heuven & L.C.W. Pols (Eds.), *Analysis and synthesis of speech: Strategic research towards high-quality TTS generation.* Mouton de Gruyter, 115-130.

Reyelt, M. (1995). 'Consistency of prosodic transcriptions. Labeling experiments with trained and untrained transcribers', *Proc. ICPhS'95*, Stockholm, 4, 290-299.

Rietveld, A.C.M. (1983). *Syllaben, klemtonen en de automatische detectie van beklemtoonde syllaben in het Nederland,* Doctoral dissertation, University of Nijmegen.

Rietveld, A.C.M. & Gussenhoven, C.H. (1985). 'On the relation between pitch excursion size and prominence', *Journal of Phonetics,* 13, 299-308.

Ross, K. & Ostendorf M. (1996). 'Prediction of abstract prosodic labels for speech synthesis', *Computer Speech and Language,* 10, 155-185.

Rumelhart, D.E., Hinton, G.E. & Williams R.J. (1986). 'Learning internal representation by error propagation', In: *Parallel distribution processing: exploration in the microstructure of cognition,* MIT Press, Cambridge.

Rump, H.H. & Hermes, D. (1996). 'Prominence lent by rising and falling pitch movements: Testing two models' *J. Acoust. Soc. Am.*, 100 (2), 1122-1131.

Silipo, R. & Greenberg, S. (1999). 'Automatic transcription of prosodic stress for spontaneous English discourse', *Proc. ICPhS'99*, San Francisco, 3, 2351-2354.

Silverman, K., Beckman, M., Pitrelli, J., Ostendorf, M., Wightman, C., Price, P., Pierrehumbert, J. & Hirschberg, J. (1992). 'TOBI: A standard for labeling English prosody', *Proc. ICSLP'92*, Banff, 2, 981-984.

Sluijter, A.M.C. & Terken, J.M.B. (1993). 'Beyond sentence prosody: Paragraph intonation in Dutch', *Phonetica*, 50, 180-188.

Sluijter, A.M.C. (1995). *Phonetic correlates of stress and accent*, Doctoral dissertation, University of Leiden.

Sluijter, A.M.C. & Heuven, V.J.J.P. van (1996). 'Spectral balance as an acoustical correlate of linguistic stress', *J. Acoust. Soc. Am.*, 100, 2471-2485.

Son, R.J.J.H. van & Pols, L.C.W. (1997). 'The correlation between consonant identification and the amount of acoustic consonant reduction' *Proc. Eurospeech'97*, Rhodes, 4, 2135-2138.

Son, R.J.J.H. van, Streefkerk, B.M. & Pols, L.C.W. (2000). 'An acoustic profile of speech efficiency', *Proc. ICSLP'00*, Peking, 1, 97-100.

Sproat, R., Hirschberg, J. & Yarowsky, D. (1992).'A corpus-based synthesizer', *Proc. ICSLP'92*, Banff, 4, 563-566.

Stöber, K., Portele, T., Wagner, P. & Hess, W. (1999). 'Synthesis by word concatenation'. *Proc. Eurospeech'99*, Budapest, 2, 619-622.

Strangert, E. & Heldner, M. (1995). 'The labeling of prominence in Swedish by phonetically experienced transcribers', *Proc. ICPhS'95*, Stockholm, 4, 204-207.

Streefkerk, B.M. (1996). 'Prominent zinsaccent en toonhoogtebewegingen', *Report of the Institute of Phonetic Sciences of the University of Amsterdam* 131, (Master Thesis).

Streefkerk, B.M. (1996). 'Prominent sentence accent and pitch movements', *Proc. Institute of Phonetic Sciences of the University of Amsterdam* 20, 111-119.

Streefkerk, B.M. (1997). 'Acoustical correlates of prominence: A design for research', *Proc. Institute of Phonetic Sciences of the University of Amsterdam* 21, 131-142.

Streefkerk, B.M., Pols, L.C.W. & Bosch, L.F.M. ten (1997). 'Prominence in read aloud sentences, as marked by listeners and classified automatically', *Proc. Institute of Phonetic Sciences of the University of Amsterdam* 21, 101-116.

Streefkerk, B.M. & Pols, L.C.W. (1998). 'Prominence in read aloud Dutch sentences as marked by naive listeners', *Tagungsband KONVENS-98*, Frankfurt a. M., 201-205.

Streefkerk, B.M., Pols, L.C.W. & Bosch, L.F.M. ten (1998). 'Automatic detection of prominence (as defined by listeners' judgements) in read aloud Dutch sentences', *Proc. ICSLP'98*, Sydney, 3, 683-686.

Streefkerk, B.M., Pols, L.C.W. & Bosch, L.F.M. ten (1999 a). 'Towards finding optimal features of perceived prominence', *Proc. ICPhS'99*, San Francisco, 3, 1769-1772.

Streefkerk, B.M., Pols, L.C.W. & Bosch, L.F.M. ten (1999 b). 'Acoustical features as predictors for prominence in read aloud Dutch sentences used in ANN's', *Proc. Eurospeech'99*, Budapest, 1, 551-554.

Streefkerk, B.M., Pols L.C.W. & Bosch, L.F.M. ten (2001). 'Up to what level can acoustical and textual features predict prominence', *Proc. Eurospeech'01*, Aalborg, 2, 811-814.

Swerts, M. & Hirschberg, J. (1998). 'Prosody and conversation: an introduction', *Language and Speech,* 41 (3-4), 229-233.

Syrdal, A.K., Hirschberg, J., McGory, J. & Beckman, M. (2001). 'Automatic ToBI prediction and alignment to speed manual labeling of prosody', *Speech Communication,* 33, 135-151.

Syrdal, A.K. & McGory, J. (2000). 'Inter-transcriber reliability of ToBi prosodic labeling', *Proc. ICSLP*, Peking, 3, 235-238.

Taylor, P. (1993). 'Automatic recognition of intonation from F0 contours using the rise / fall / connection model', *Proc. Eurospeech'93*, Berlin, 2, 789-792.

Taylor, P., King, S., Isard, S. & Wright, H. (1998).'Intonation and dialog context as constraints for speech recognition', *Language and Speech,* 41, 493-512.

Terken, J. (1991). 'Fundamental frequency and perceived prominence of accented syllables', *J. Acoust. Soc. Am.*, 89 (4), 1768-1776.

Terken, J. & Hirschberg, J. (1994). 'Deaccentuation of 'words', representing given information: effects of persistence of grammatical function and surface position', *Language and Speech,* 37 (2), 125-145.

Terken, J. (1995). 'The generation of prosodic structure and intonation in speech synthesis', In: W.B. Kleijn & K.K. Paliwal (Eds.), *Speech coding and synthesis*, 635-662.

Terken, J. (1996). 'Variation of accent prominence within the phrase: Models and spontaneous speech data', In: Y. Sagisaka, N. Campbell & N. Higuchi (Eds.), *Computing Prosody. Approaches to a Computational Analysis and Modelling of the Prosody of Spontaneous Speech*, Springer-Verlag, New York, 95-116.

Tournemire, S. de (1998). 'Automatic transcription of intonation using an identified prosodic alphabet', *Proc. ICSLP'98*, Sydney, 5, 1955-1958.

Vereecken, H., Martens, J-P., Grover, C., Fackrell, J. & Coile, B. van (1998). 'Automatic labeling of 6 languages', *Proc. ICSLP'98,* Sydney, 4, 1399-1402.

Vaissière, J. (1989). 'On the automatic extraction of prosodic information for automatic speech recognition system', *Proc. Eurospeech'89*, Paris, 1, 202-205.

Véronis, J., Di Cristo, P., Courtois, F. & Lagrue, B. (1997). 'A stochastic model of intonation for French text-to-speech synthesis', *Proc. Eurospeech'97*, Rhodes, 5, 2643-2646.

Véronis, J. & Campione, E. (1998). 'Towards a reversible symbolic coding of intonation', *Proc. ICSLP'98*, Sydney, 7, 2899-2902.

Véronis, J., Di Cristo, P., Courtois, F, & Chaumette, C. (1998). 'A stochastic model of intonation for text-to-speech synthesis', *Speech Communication,* 26, 233-244.

Wang, C. (2001). *Prosodic modeling for improved speech recognition and understanding*, Doctoral dissertation, MIT.

Wang, C. & Seneff, S. (2001). 'Lexical stress modeling for improved speech recognition of spontaneous telephone speech in the Jupiter domain', *Proc. Eurospeech'01*, Aalborg, 3, 2761-2764.

Wang, X. (1997). *Incorporating knowledge on segmental duration in HMM-based continuous speech recognition*, Doctoral dissertation, University of Amsterdam.

Weenink, D.J.M. (1991). 'Aspects of neural nets' *Proc. Institute of Phonetic Sciences of the University of Amsterdam*, 15, 1-25

Widera, C., Portele, T. & Wolters, M. (1997). 'Prediction of word prominence' *Proc. Eurospeech'97*, Rhodes, 2, 999-1002.

Wightman, C.W., Shattuck-Hufnagel, S., Ostendorf, M. & Price, P.J. (1992). 'Segmental durations in the vicinity of prosodic phrase boundaries', *J. Acoust. Soc. Am.*, 91 (3), 1707-1717.

Wightman, C.W. & Ostendorf, M. (1994). 'Automatic labeling of prosodic patterns', *Proc. IEEE'94*, 2/4, 469-481.

Wightman, C. W. & Rose, R. C. (1999). 'Evaluation of an efficient prosody labeling system for spontaneous speech utterances' *Proc. IEEE Automatic Speech Recognition and Understanding Workshop*, Keystone.

Wightman, C.W., Syrdal, A.K., Stemmer, G., Conkie, A. & Beutnagel, M. (2000). 'Perceptually based automatic prosody labeling and prosodically enriched unit selection improve concatenative text-to-speech synthesis', *Proc. ICSLP'00*, Peking, 2, 71-74.

Wright, H. & Taylor, P. (1997). 'Modelling intonational structure using hidden Markov models', *ESCA Workshop on Intonation Theory, Models and Applications*, Athens, 333-336.

Wijk, C. van & Kempen, G. (1979). 'Functiewoorden; een inventatrisatie voor het Nederlands', *Intern rapport*, 79 FU 05.

SUMMARY

The purpose of this study was to explore the notion of prominence in spoken language. It concentrated on finding an operational definition of prominence, on giving a description of the linguistic and acoustical correlates of prominence, and on analyzing these correlates in terms of their contribution to prominence distinctions. Furthermore, this study was concerned with feature extraction, and with prominence prediction, either on the basis of linguistic features or on the basis of acoustic features.

In **chapter 1** the notion of prominence is explained, the use of prominence in (speech) communication is illustrated and research questions are described.
In speech some parts are more prominent than others. This is a gradient property. In many languages prominence helps to structure the message e.g. the prominent parts are the important ones. In addition, prominence helps to increase the comprehensibility and the naturalness of speech.
The listener uses two information sources to perceive prominence levels: bottom-up information and top-down information. The listener uses cues from the speech signal such as speech segments being louder, being longer and being realized with a pitch movement (bottom-up information) to detect prominence. The expectation of prominence is built on the basis of his / her knowledge of the language (top-down information).
From a phonetic viewpoint prominence is closely related to the notion of pitch accent and lexical (word) stress. Prominence is a perceptual phenomenon and is intuitively clear to non-experts. Prominence can function as an interface between acoustics and aspects of structure e.g. in terms of 'given' and 'new' information. The prediction of prominence may also be useful in speech technology.
The research questions concentrate mainly on the following: 1) how to find an operational definition of prominence, 2) which are the linguistic determinants / correlates of prominence, and 3) which acoustic correlates can be found. The implementation part of this research concentrates on the automatic extraction of features, on the analysis, and on the prediction of prominence on the basis of the preselected features.

In **chapter 2** a perceptual definition of prominence is investigated. The read-aloud sentences of the Dutch Polyphone Corpus (telephone speech) are used as research

material, which unavoidably contains a great deal of speaker variability, and which is typical for many speech-technology applications.
The prominence-marking task was made as easy as possible to the subjects, giving them as much freedom for their own interpretation of prominence as possible, and allowing listeners to label large amounts of data. It was decided to mark prominence in a binary rather than multi-valued way, because otherwise the task was too time consuming and multi-valued marks from each listener appeared not to be necessary since it was shown that the cumulative marks also provide gradient prominence information. However, the results of a pilot-experiment were not very convincing, it was concluded that listeners mark prominence at the word level more consistently than at the syllable level. Since the unit of a word is also more meaningful to naive listeners than syllables, the word was chosen as the unit to mark.

In this research prominence was made operational in the following way: ten listeners were asked to mark those word(s) that they considered were spoken with emphasis. The cumulative marks of listeners provided detailed information about the degree of prominence of each word. In such a way the 1244 sentences of the training set were marked for prominence. One 'optimal' listener, just giving binary judgments only, marked the independent test set of another 1000 sentences. This relatively simple binary marking allowed for an annotation of word prominence for more than 4.5 hours of speech. The listeners were rather consistent (mean agreement expressed in Cohen's Kappa $\kappa = 0.50$) and reliable. Many of the inconsistencies could be attributed to shifts of the individual prominence detection thresholds. However, threshold shifts and differences occur, which influence the agreement measure negatively.

In **chapter 3** linguistic correlates of prominence are described, analyzed and used as predictors for prominence.
Relationships between, on the on hand, (1) Part-of-Speech (e.g. Noun, Adverb, Article), (2) word length, (3) position of a word in the sentence, and (4) interdependency of Part-of-Speech categories such as Adjective-Noun combinations and, on the other hand, prominence are described and analyzed in detail. Word classes are ranked according to increasing prominence and word length appears to be related to prominence. In generally, the longer the words the more prominent they are. Nouns occurring in Adjective-Noun combinations tend to be less prominent than in all other combinations and the first content word in a sentence is more prominent than the content words occurring at other positions in the sentence.
Based on these relationships an algorithm was developed to predict prominence degrees. This gradient prominence prediction, especially in the middle part of the scale, is more problematic. However, the reduced binary prominence prediction is correct in 81% of the cases for the independent test set.
Concluding, one is able to select a simple set of automatically derived linguistic features, which predicts prominence with the same agreement as listeners do

(Cohen's Kappa $\kappa = 0.62$), indicating that top-down information can provide enough to predict prominence accurately. However, some used linguistic relationships may be specific for this type of speech / text material.

Chapter 4 deals with the description and detailed analysis of the acoustic features of prominence. It concentrates on the feature extraction on the level of the individual word and does not take the neighboring words into account. This research suggests that the following features are useful for predicting prominence: (1) F_0 range per word, (2) F_0 range per syllable, (3) syllable duration, (4) vowel intensity, (5) median F_0 per syllable, (6) vowel duration, (7) normalized vowel intensity and (8) normalized vowel duration. The above order gives also the ranking with respect to the features' ability to discriminate between prominent and non-prominent words:
It was striking that when the vowels were normalized for their intrinsic properties, such as intrinsic vowel duration and intrinsic vowel intensity, the discrimination was no better than using the unnormalized counterparts. It was hypothesized that the variability in the speech material used was too large to properly correct for intrinsic durational and other properties.

Chapter 5 deals with the question whether the selected set of analyzed acoustic correlates could be used as input features in order to recognize prominence. It was shown that apart from F_0, syllable duration (more than vowel duration) and vowel intensity were useful input features for a recognition device. Automatic extraction of acoustic features was performed in such a way that a binary neural net classification resulted in a best recognition rate of 79% correct on the independent test set. The agreement of the predicted prominence (Cohen's Kappa $\kappa = 0.50$) was at least as good as the mean agreement of the listeners. This result was achieved with only twelve input features. Gradient prominence prediction on a 10-point scale is more difficult and requires further research.

In the last chapter (**chapter 6**) all findings and conclusions are summarized. Naive listeners were able to mark prominent words in spoken sentences with some consistency and reliability. The results of this study showed that acoustic and linguistic correlates of prominence can be determined automatically and they can be used to predict prominence either on text or on the speech signal.
Prominence assignment of naive listeners is valuable because the determined acoustic correlates, related to bottom-up information, and linguistic correlates, related to top-down information, describe the perceptual notion of prominence. This research shows that the prediction of prominence by acoustic or linguistic features is undistinguishable from prominence assigned by naive listeners.

SAMENVATTING

In gesproken taal worden bepaalde woorden als nadrukkelijker (prominenter) waargenomen dan andere. Prominentie (nadruk) is voor de communicatie van groot belang omdat prominentieverschillen o.a. structuur geven aan de boodschap. De belangrijkste vraag in dit onderzoek is: welke kenmerken in gesproken taal leiden ertoe dat bepaalde woorden in een zin prominenter worden waargenomen dan andere woorden. Oftewel, waardoor horen we of weten we dat een woord met nadruk is gezegd?
Aan de orde komen de gebruikte operationele definitie van prominentie, het beschrijven van linguïstische en akoestische correlaten van prominentie, en het analyseren in hoeverre deze correlaten een onderscheid kunnen maken tussen bijvoorbeeld prominente en niet prominente woorden. Dit onderzoek besteedt ook aandacht aan automatische kenmerkextractie en voorspelling van prominentie met behulp van akoestische of linguïstische kenmerken.

In **hoofdstuk 1** wordt het begrip prominentie uitgelegd, komt het gebruik van prominentie in gesproken taal aan de orde en worden de onderzoeksvragen beschreven.
Zoals al eerder gezegd worden sommige spraakfragmenten in gesproken taal als prominenter waargenomen dan andere. Dit is een graduele eigenschap en geen binaire. In veel talen structureert het verschil in prominentie de boodschap van de spreker in belangrijke en minder belangrijke delen. Deze prominentieverschillen maken de boodschap begrijpelijk en verhogen tevens de verstaanbaarheid en de natuurlijkheid. In dit proefschrift wordt prominentie als perceptueel fenomeen gehanteerd, hoewel het gerelateerd is aan concepten zoals toonhoogte-accent en woordklemtoon.
Voor de interpretatie van het spraaksignaal gebruikt de luisteraar twee soorten informatiebronnen; de informatie in het spraaksignaal (bottom-up) en zijn kennis van de taal (top-down).
In het spraaksignaal kunnen spraakeenheden zoals klinkers en / of lettergrepen langer van duur zijn, luider worden uitgesproken, en / of door middel van een toonhoogtebeweging prominent gemaakt worden. Hoe deze akoestische kenmerken uit het signaal te extraheren zijn en wat hun relatie is tot waargenomen prominentie, wordt nader onderzocht in dit proefschrift.
Daarnaast heeft de luisteraar door middel van zijn kennis van de taal een verwachting welke woorden prominent zullen zijn. Deze prominentieverwachting hangt nauw samen met lexicaal, syntactische, pragmatische en semantische aspecten

van de boodschap. De onderzoeksvraag in dit onderzoek heeft met deze linguïstische kenmerken van prominentie te maken: welke kenmerken kunnen automatisch geëxtraheerd en gebruikt worden om prominentie te voorspellen? Automatisch voorspellen van prominentie kan van nut zijn in de spraaktechnologie, aangezien prominentie toegang kan geven tot de informatiestructuur in termen als 'gegeven' of 'nieuwe' informatie.

In **hoofdstuk 2** wordt prominentie operationeel gedefinieerd en wordt ingegaan op de consistentie en de betrouwbaarheid van de luisteraars die de prominentie-oordelen afgeven. Als onderzoeksmateriaal zijn los voorgelezen zinnen uit het Nederlandse Polyphone Corpus (telefoonspraak) gebruikt. Dit materiaal heeft veel sprekervariabiliteit. Bij het formuleren van de prominentie-markeer-taak is geprobeerd de luisteraar zo veel mogelijk vrijheid te geven zijn eigen interpretatie van het begrip prominentie toe te passen (de taak voor de luisteraar was om woorden te markeren zodra die volgens de luisteraar met nadruk waren uitgesproken). Verder is de taak zo simpel mogelijk gehouden, door de luisteraar een binair en geen gradueel prominentie-oordeel te laten geven. De gesommeerde prominentie-oordelen van meerdere luisteraars geven ook inzicht in de verschillende prominentiegraden. Er is besloten om de luisteraars prominentie op woorden en niet op lettergrepen te laten markeren, omdat een pilot erop wees dat deze taak consistentere oordelen zou opleveren. Bovendien heeft de eenheid van een woord voor de naïeve luisteraar een duidelijkere betekenis dan de eenheid van een lettergreep. Deze aanpak maakte de markeertaak valide.
Het resultaat was dat het trainingsmateriaal van 1244 zinnen door tien luisteraars voor prominentie gemarkeerd is. De gesommeerde binaire markeringen geven per woord een graduele markering op een schaal van 0 tot 10. De luisteraars vertoonden verschil in 'gevoeligheid' omtrent het markeren van prominentie. Sommige luisteraars markeerden duidelijk meer woorden in een zin dan andere. Ondanks dit negatieve effect op de mate van overeenstemming ligt de gemiddelde overeenstemming, uitgedrukt in Cohen's Kappa, bij 0.50. Het markeren van de onafhankelijke testset van 1000 zinnen is verder versimpeld door slechts één luisteraar (de meest consistente en betrouwbare) deze set te laten markeren. Deze relatief eenvoudige taken maakten het mogelijk dat in totaal 4,5 uur aan spraak op woordniveau gemarkeerd is voor prominentie.

In **hoofdstuk 3** worden linguïstische kenmerken van prominentie beschreven en geanalyseerd die daarna tevens gebruikt worden om prominentie automatisch te voorspellen.
Er is beschreven dat woordklassen (zelfstandig naamwoord, voornaamwoord, lidwoord enz.) gerangschikt kunnen worden op toenemende prominentie. Ook woordlengte, uitgedrukt in het aantal lettergrepen per woord, hangt nauw samen met prominentie. Des te langer een woord des te prominenter. Een zelfstandig naamwoord voorafgegaan door een bijvoeglijk naamwoord is vaak minder prominent dan in een andere combinatie. Het eerste inhoudswoord in de zin is vaak prominenter dan inhoudswoorden op andere posities.

Deze relaties tussen prominentie enerzijds en lexicale en syntactische kenmerken anderzijds worden gebruikt om met behulp van een eenvoudige 'heuristische' regelset prominentie in vijf gradaties te voorspellen. Voorspelling van prominentie in de middengradaties levert problemen op; reductie van deze schaal tot een binaire geeft echter 81% overeenstemming wanneer de voorspelde prominentie wordt vergeleken met de binaire prominentiemarkering van de testset. De mate van overeenstemming van de voorspelde prominentie met handgemarkeerde prominentie ligt bij dezelfde Kappa-waarden als de overeenstemming tussen luisteraars.
Hieruit blijkt dat linguïstische kenmerken in staat zijn prominentie op een adequate manier te beschrijven en dat de automatische voorspelling ononderscheidbaar is van die van naïeve luisteraars.

Hoofdstuk 4 beschrijft gedetailleerde analyses van akoestische correlaten van prominentie. De nadruk ligt op kenmerkextractie uit de individuele prominente woorden en de bijbehorende lettergrepen en klinkers, en niet op kenmerkextractie gerelateerd aan de woorden in de omgeving, noch op de intonatiecontour zelf. Een gedetailleerde analyse leverde de hieronder genoemde bruikbare correlaten op. Door middel van een analyse konden deze correlaten gerangschikt worden met betrekking tot hun onderscheidingsvermogen tussen prominente en niet prominente woorden: (1) F_0-range[1] per woord, (2) F_0-range per lettergreep, (3) duur van de lettergreep, (4) klinkerintensiteit, (5) mediaan F_0 per lettergreep, (6) duur van de klinker, (7) genormaliseerde klinkerintensiteit en (8) genormaliseerde duur van de klinker. Het valt op dat de normalisaties voor de intrinsieke eigenschappen, zoals b.v. het feit dat iedere klinker zijn eigen duur heeft, niet (systematisch) leiden tot een beter onderscheidingsvermogen. Verder onderzoek is nodig om uit te zoeken of het de grote sprekervariabiliteit of de automatische segmentatie is die de normalisaties beïnvloedt, of dat andere factoren een rol spelen.

Hoofdstuk 5 concentreert zich op de vraag, of de in hoofdstuk 4 geselecteerde correlaten gebruikt kunnen worden als inputkenmerken voor herkenning van prominentie met behulp van een neuraal netwerk. Bij herkenning met neurale netwerken bleek dat naast F_0-range per woord ook de lettergreepduur en de klinkerintensiteit bruikbare kenmerken zijn voor prominentieherkenning, beter zelfs dan klinkerduur. Twaalf automatisch geëxtraheerde kenmerken zijn gebruikt, zodat een binaire classificatie (prominent of niet prominent) voor 79% overeenstemt met de markeringen van de luisteraar op de testset. Uitgedrukt in de overeenstemmingsmaat Cohen's Kappa was dat $\kappa = 0.50$. Dit betekent dat prominentiemarkeringen door het neurale netwerk wat betreft consistentie niet te onderscheiden zijn van die van luisteraars. Graduele prominentieherkenning met name in de middencategorieën blijkt iets moeilijker te zijn en vergt nog verder onderzoek.

In **hoofdstuk 6** worden alle bevindingen en conclusies samengevat. Luisteraars zijn in staat met enige consistentie en betrouwbaarheid prominente woorden in zinnen te

[1] ook vertaald met spreidingsbreedte

markeren. Het perceptuele fenomeen van prominentie is daarmee operationeel geformuleerd. De resultaten van dit onderzoek laten zien dat zowel de akoestische alsook de linguïstische correlaten bepaald zijn die prominentie op een adequate manier beschrijven en kunnen voorspellen. Prominentie gemarkeerd door naïeve luisteraars is waardevol omdat zowel lexicaal / syntactische kenmerken, die nauw samenhangen met top-down informatie, als ook akoestische kenmerken, die nauw samenhangen met bottom-up informatie, het perceptuele begrip prominentie voldoende beschrijven. De voorspelling van prominentie is zowel op grond van lexicaal / syntactische inputkenmerken alsook op grond van akoestische kenmerken niet te onderscheiden van die van naïeve luisteraars.

CURRICULUM VITAE

Barbertje Streefkerk was born in Amsterdam on 18 August 1972. She attended the primary school in Gouda at the Vrije School e.o. from 1978 to 1981. From 1981 to 1992 she attended the Freie Waldorfschule in Gladbeck (Germany), where she obtained her Abitur in 1992. From 1992 to 1996 she studied German literature and linguistics, and alfa-informatica (phonetics) at the University of Amsterdam. During her study she spent four month at the Institut für Kommunikationsforschung und Phonetik at the Rheinische Friedrich-Wilhelms Universität Bonn. In December 1996 she obtained her Master's degree in phonetic sciences, with a specialization in speech technology. From 1997 to 2001 she was employed as a PhD student by the Amsterdam Center for Language and Communication (ACLC). The research presented in this PhD thesis was carried out during that period at the Institute of Phonetic Sciences Amsterdam (IFA).

Barbertje Streefkerk

Prominence
Acoustic and lexical/syntactic correlates

In spoken language some words are perceived as more prominent than others. Without these differences in prominence spoken language is unclear and boring.

In this PhD thesis both acoustic and lexical / syntactic correlates of perceived prominence are discussed. *Prominence* is defined at the word level and naive listener judgements are used as the norm. It is related both to pitch accent and lexical (word) stress. One of the findings in this thesis is that naive listeners are able to mark word prominence rather consistently on isolated Dutch sentences.

A selected set of *acoustic* input features is used for classification of prominent words using feed-forward neural networks. On the basis of an optimally selected set, we obtained an accuracy of 79% in prominence classification on a test set containing 1000 sentences.
Using *lexical / syntactic* input features (such as word class, word length and position of the word in the sentence), which are derived from text only, an algorithm to predict prominence is developed. The predicted prominence agrees with the perceived prominence in 81% of the cases for the test set.

The results show that acoustic and linguistic correlates of prominence can be determined automatically and can be used to accurately predict prominence. Statistical agreement measures show that prominence prediction on the acoustic as well as on the lexical / syntactic input level is undistinguishable from prominence assignment by naive listeners. For phonetics this PhD thesis gives insight into the human recognition process of prominence. For speech technology, knowing the prominent and non-prominent words may be useful, for instance, to disambiguate the meaning of two similar sentences.

This book is of interest for researchers in the fields of phonetics, prosody and speech technology.

Barbertje Streefkerk (1972) studied German literature and linguistics, and phonetic sciences (technology) at the University of Amsterdam from 1992 to 1996. From 1997 to 2001 she was employed as a PhD student by the Amsterdam Center for Language and Communication (ACLC). The research presented in this PhD thesis was carried out during that period at the Institute of Phonetic Sciences Amsterdam (IFA).

: LOT
Netherlands
Graduate
School of
Linguistics

ISBN 90-76864-19-5